www.sapientia.uji.es | 41



Métodos matemáticos

José Antonio López Ortí

Métodos matemáticos

José Antonio López Ortí



Departamento de Matemáticas

■ Codi d'assignatura 321

Edita: Publicacions de la Universitat Jaume I. Servei de Comunicació i Publicacions
Campus del Riu Sec. Edifici Rectorat i Serveis Centrals. 12071 Castelló de la Plana
http://www.tenda.uji.es e-mail: publicacions@uji.es

Col·lecció Sapientia, 41
Primera edició, 2010
www.sapientia.uji.es

ISBN: 978-84-693-4783-6



Índice General

Tema 1

Teoría de grafos

1.1. Introducción

En este tema se aborda el estudio de la teoría de grafos, así como sus principales aplicaciones a problemas de optimización. Para ello, en la primera parte del mismo, se introducirá la definición de grafo simple y de dígrafo; a continuación se introducirán estructuras de datos adecuadas para la representación de éstos. También en esta primera parte se introducirán los distintos tipos de recorridos en grafos, así como el concepto de conexión; a continuación se introducirán los conceptos de grafos euleriano y hamiltoniano. El último tópico que se abordará en esta primera parte lo constituye el concepto de árbol, el cuál se definirá y caracterizará.

La segunda parte de este primer tema se dedicará al estudio de los grafos y dígrafos ponderados, abordándose el estudio del problema del camino más corto entre dos puntos de un grafo ponderado, el problema del conector mínimo, el problema de la trayectoria crítica y el problema del flujo máximo en una red.

1.2. Grafos y dígrafos

Antes de comenzar con las definiciones, se ha considerado importante advertir que, en teoría de grafos, desgraciadamente no existe una notación unificada, por lo que a la hora de realizar un estudio comparativo de diversos textos es de suma importancia establecer con claridad qué definiciones y conceptos utiliza cada uno de ellos, pues de otro modo no sería posible realizar dicho estudio comparado.

1.2.1. Definiciones generales

Definición 1.2.1
Se define **grafo simple** G como un par $(V(G), E(G))$ donde $V(G)$ es un conjunto finito no vacío, $V(G) = \{v_1, v_2, \ldots, v_n\}$, a cuyos elementos v_i llamamos vértices (o nodos) y $E(G)$ es un conjunto cuyos elementos son subconjuntos de dos elementos de $V(G)$ a los cuales llamamos aristas; así $E(G) = \{\{u, v\}; u, v \in V(G), u \neq v\}$.

A continuación se introduce una serie de conceptos muy usuales en teoría de grafos:

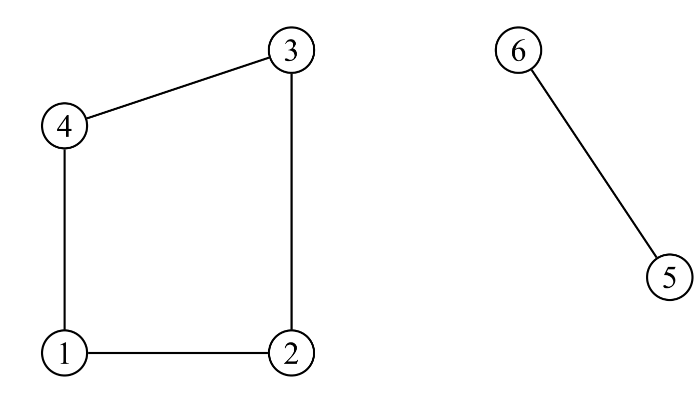
- Dada una arista $\{u, v\}$, llamamos extremos de la arista a los vértices u, v.
- Diremos que una arista e incide en un vértice v si tiene a éste por extremo: esto es $e = \{v, w\}$ con $w \in V(G)$.
- Diremos que dos vértices u, v de un grafo G son adyacentes si $\{u, v\} \in E(G)$.
- Diremos que dos aristas son adyacentes si tienen un vértice común.
- Diremos que una arista $e \in E(G)$ conecta los vértices u, $v \in E(G)$ si $e = \{u, v\}$.
- Llamamos orden, grado, o valencia de un vértice $v \in V(G)$ al número de aristas distintas que contienen a v.
- Llamamos vértice aislado a aquél que tiene grado cero.
- Llamamos vértice terminal a aquél que tiene grado uno.

En teoría de grafos es frecuente utilizar la notación $V(G) = \{v_1, \ldots, v_n\}$ para indicar el conjunto de vértices de un grafo y $E(G) = \{\{v_{i_1}, v_{j_1}\}, \ldots, \{v_{i_r}, v_{j_s}\}\}$ para indicar el conjunto de aristas. También es corriente representar los vértices como $V(G) = \{1, 2, \ldots, n\}$. Una característica de los grafos es su representabilidad, así se tiene:

Ejemplo 1.2.1
Sea el grafo G cuyos sus vértices y aristas vienen dados respectivamente por:

$$V(G) = \{1, 2, 3, 4, 5, 6\}$$

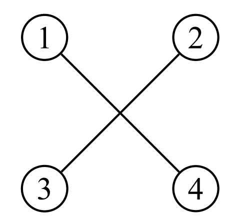
$$E(G) = \{\{1, 2\}, \{2, 3\}, \{3, 4\}, \{1, 4\}, \{5, 6\}\}$$



La figura anterior constituye la representación gráfica del grafo G definido anteriormente. ♦

Cuando se representan grafos debe ser tenido en cuenta que dos aristas únicamente pueden incidir en un vértice, por lo que representaciones como la siguiente no implican que las aristas se corten.

Ejemplo 1.2.2
Las aristas son $\{1, 4\}$ y $\{2, 3\}$.



♦

Definición 1.2.2
Diremos que dos **grafos** $G_1 = (V(G_1), E(G_1))$ y $G_2 = (V(G_2), E(G_2))$ son **isomorfos** si $\exists\phi : V(G_1) \to V(G_2)$ biyectiva de modo que $\{v, w\} \in E(G_1)$ si y sólo si $\{\phi(v), \phi(w)\} \in E(G_2)$.

Según esta definición, podemos decir de un modo más intuitivo que dos grafos son isomorfos si es posible renombrar los vértices de uno de ellos de modo que ambos grafos resulten coincidentes.

Ejemplo 1.2.3
Sean G, G' los grafos abajo representados a izquierda y derecha respectivamente



Ambos grafos son isomorfos, pues es posible establecer entre ellos la aplicación $\psi : V(G) \to V(G')$ dada por $\psi(1) = a$, $\psi(2) = d$, $\psi(3) = c$, $\psi(4) = b$ la cual es biyectiva y cumple los requerimientos de la definición 1.2.2. ♦

A continuación se introduce el concepto de subgrafo como:

Definición 1.2.3
Diremos que un grafo $S = (V(S), E(S))$ es un **subgrafo** de un grafo $G = (V(G), E(G))$ si $V(S) \subset V(G)$ y $E(S) \subset E(G)$.

Ejemplo 1.2.4
Sea G el grafo dado en el ejemplo 1.2.1, y sea S el grafo cuyos conjunto de vértices y aristas vienen dados respectivamente por $V(S) = \{2, 3, 4, 5, 6\}$ y $E(S) = \{\{2, 3\}, \{5, 6\}\}$.

El grafo S es subgrafo de G pues, por una parte, se verifica que

$$V(S) = \{2, 3, 4, 5, 6\} \subset \{1, 2, 3, 4, 5, 6\} = V(G)$$

y por otra

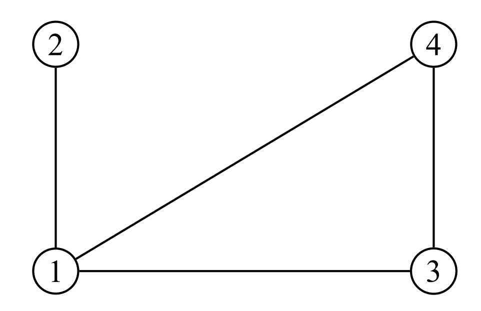
$$E(G) = \{\{2,3\},\{5,6\}\} \subset \{\{1,2\},\{2,3\},\{3,4\},\{1,4\},\{5,6\}\}$$

♦

A continuación se procede a dar una primera definición de grafo conexo, concepto que se precisará mejor más adelante.

Definición 1.2.4
Sea G un grafo; sea $V(G) = \{1, 2, \ldots, n\}$ su conjunto de vértices. Diremos que G es **conexo** si para todo par de vértices $i, j \in V(G)$ existe un conjunto de aristas $\{e_1, \ldots, e_k\} \in E(G)$ de modo que si $e_l = \{u_l, v_l\}$ se satisface $v_l = u_{l+1}$, $l = 1, \ldots, k-1$, $u_1 = i$, $v_k = j$.

Ejemplo 1.2.5
Sea el grafo representado en la figura



Dicho grafo es conexo, pues para los posibles pares de vértices se verifica:

- 1, 2 los cuales pueden conectarse mediante $\{1, 2\}$.
- 1, 3 los cuales pueden conectarse mediante $\{1, 3\}$.
- 1, 4 los cuales pueden conectarse mediante $\{1, 4\}$.
- 2, 3 los cuales pueden conectarse mediante $\{2, 1\}$, $\{1, 3\}$.
- 2, 4 los cuales pueden conectarse mediante $\{2, 1\}$, $\{1, 4\}$.
- 3, 4 los cuales pueden conectarse mediante $\{3, 4\}$.

Por tanto el grafo es conexo.

Como ejemplo de grafo no conexo podemos citar el que aparece en el ejemplo 1.2.1, pues en él no es posible conectar mediante aristas adyacentes los vértices 1 y 5. ♦

1.2.2. Matriz de adyacencia y matriz de incidencia

A continuación se procede a introducir estructuras que permitan la representación algebraica de grafos, de modo que éstos, a través de las mismas, puedan ser introducidos de modo eficiente en distintos algoritmos. Esto es necesario para poder proceder al manejo computacional.

Definición 1.2.5
Sea G un grafo simple, cuyo conjunto de vértices posee n elementos. Llamamos **matriz de adyacencia** de G a la matriz $n \times n$, $Ad(G)$, definida como

$$Ad(G) = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \dots & \dots & \dots & \dots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{bmatrix} \qquad \text{donde} \qquad a_{i,j} = \begin{cases} 1, & \text{si } \{i,j\} \in E(G) \\ 0 & \text{si } \{i,j\} \notin E(G) \end{cases}$$

Nótese que la matriz $Ad(G)$ de un grafo simple es simétrica.

Ejemplo 1.2.6
En el caso del grafo del ejemplo 1.2.1 se tiene que la **matriz de adyacencia** viene dada por

$$Ad(G) = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

♦

Definición 1.2.6
Sea un grafo simple $G = (V(G), E(G))$ y sean n, m los cardinales de $V(G) = \{1, 2, \dots, n\}$ y de $E(G) = \{e_1, \dots, e_m\}$, respectivamente. Llamamos **matriz de incidencia** del grafo G, a la matriz $n \times m$, $In(G)$, definida como

$$In(G) = \begin{bmatrix} b_{1,1} & b_{1,2} & \dots & b_{1,m} \\ b_{2,1} & b_{2,2} & \dots & b_{2,n} \\ \dots & \dots & \dots & \dots \\ n_{n,1} & b_{n,2} & \dots & b_{n,m} \end{bmatrix} \qquad \text{donde,} \qquad b_{i,j} = \begin{cases} 1 & \exists\, k \in V(G) : \, e_j = \{i,k\} \\ 0 & \nexists k \in V(G) : \, e_j = \{i,k\} \end{cases}$$

Ejemplo 1.2.7
Para el grafo anteriormente representado, ordenando las aristas como

$$E(G) = \{\{1,2\}, \{1,4\}, \{2,3\}, \{3,4\}, \{5,6\}\}$$

la matriz de incidencia resulta:

$$In(G) = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

♦

1.2.3. Operaciones con grafos

A continuación se definen algunas de las operaciones más comunes que se pueden encontrar en teoría de grafos, esto es, la unión y suma de grafos, las operaciones de supresión de vértices y aristas, la contracción según una arista, y la construcción del grafo dual. Así se tiene:

Definición 1.2.7
Sean G_1, G_2 dos grafos, llamamos **unión de los grafos** G_1 y G_2 al grafo $G_1 \cup G_2$ definido por $V(G_1 \cup G_2) = V(G_1) \cup V(G_2)$, $E(G_1 \cup G_2) = E(G_1) \cup E(G_2)$.

Ejemplo 1.2.8
Sean los grafos G y G' dados respectivamente por $V(G) = \{1,2,3,4\}$, $E(G) = \{\{1,2\},\{2,3\},\{3,4\},\{1,4\}\}$ y $V(G') = \{5,6\}$, $E(G') = \{\{5,6\}\}$. La unión de los grafos es el grafo $G \cup G'$ definido por $V(G \cup G') = V(G_1) \cup V(G_2) = \{1,2,3,4,5,6\}, E(G \cup G') = E(G) \cup E(G_2) = \{\{1,2\},\{2,3\},\{3,4\},\{1,4\},\{5,6\}\}$.
♦

Si $Ad(G)$ y $Ad(G')$ son las matrices de adyacencia de los grafos G, G', la matriz de adyacencia de la unión la podemos representar en el caso $V(G_1) \cap V(G_2) = \emptyset$ como:

$$Ad(G \cup G') = \begin{bmatrix} Ad(G) & 0 \\ 0 & Ad(G') \end{bmatrix}$$

Así, en el ejemplo anterior resulta:

$$Ad(G) = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}, \ Ad(G') = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \ Ad(G \cup G') = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

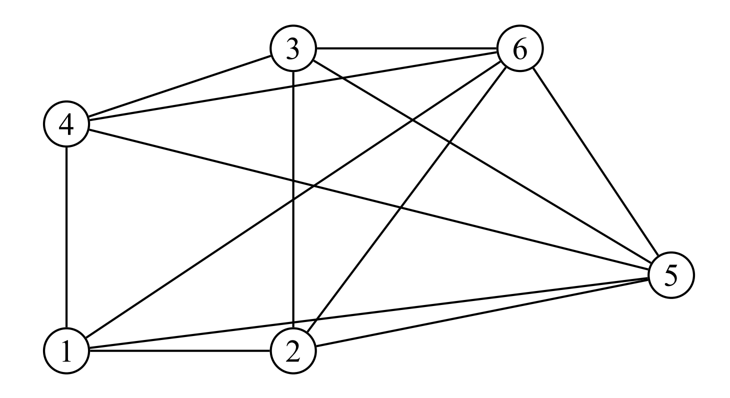
Definición 1.2.8
Sean G_1, G_2 dos grafos, llamamos **suma de los grafos** G_1 y G_2 al grafo $G_1 + G_2$ definido como $V(G_1 + G_2) = V(G_1)) \cup V(G_2)$, $E(G_1 + G_2) = E(G_1) \cup E(G_2) \cup \{\{u,v\} | u \in V(G_1), v \in V(G_2)\}$.

Ejemplo 1.2.9
Consideremos los grafos del ejemplo anterior; el grafo suma $G + G'$ donde

$$V(G + G') = \{1,2,3,4,5,6\}$$

$$E(G + G') = \{\{1,2\},\{2,3\},\{3,4\},\{1,4\},\{5,6\},\{1,5\},\{1,6\}, \\ ,\{2,5\},\{2,6\},\{3,5\},\{3,6\},\{4,5\},\{4,6\}\}$$

y cuya representación gráfica viene dada por:



♦

La matriz de adyacencia de la suma de grafos cuando $V(G_1) \cap V(G_2) = \emptyset$ viene dada por

$$Ad(G_1 + G_2) = \begin{bmatrix} Ad(G_1) & 1_{n_1} \\ 1_{n_2} & Ad(G_2) \end{bmatrix}$$

donde 1_k representa la matriz $k \times k$ cuyos elementos son todos 1; n_1 el cardinal de $V(G_1)$ y n_2 el de $V(G_2)$. Así, para el ejemplo anterior se tiene

$$Ad(G) = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}, \quad Ad(G') = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad Ad(G+G') = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

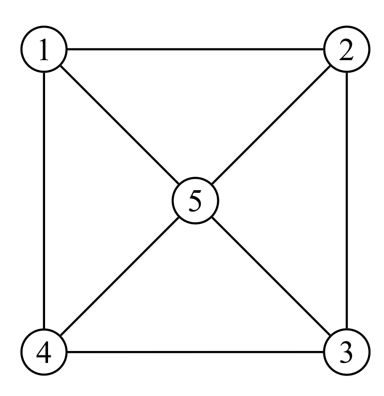
Definición 1.2.9
Sea G un grafo simple, y sea $v \in V(G)$. Llamamos **supresión en G del vértice v** al grafo $G - v$ definido como $V(G - v) = V(G) - \{v\}$, $E(G - v) = \{e \in E(G) | e \neq \{v, w\}\}$.

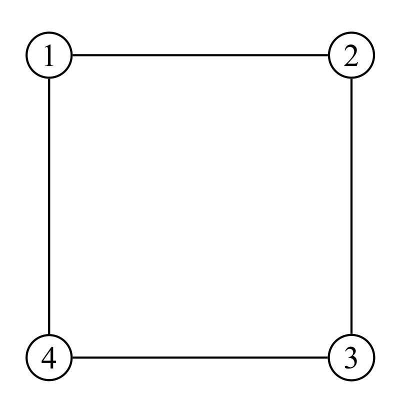
Ejemplo 1.2.10
Sea el grafo G definido por $V(G) = \{1, 2, 3, 4, 5\}$

$$E(G) = \{\{1, 2\}, \{1, 4\}, \{1, 5\}, \{2, 3\}, \{2, 5\}, \{3, 4\}, \{3, 5\}, \{4, 5\}\}$$

y sea el vértice $5 \in V(G)$, el grafo $G - 5$ tiene como conjunto de vértices $V(G - 5) = \{1, 2, 3, 4\}$ y como conjunto de aristas $E(G - 5) = \{\{1, 2\}, \{1, 4\}, \{2, 3\}, \{\{3, 4\}\}$.

En la siguiente representación la figura que aparece a la izquierda representa G, siendo la que aparece a la derecha una representación de $G - 5$.



,


♦

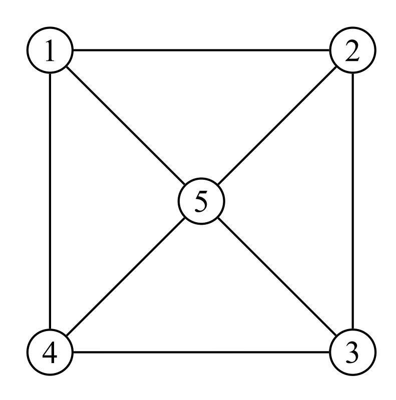
La matriz de adyacencia correspondiente a la supresión en G de un vértice i es la matriz menor complementaria de $Ad(G)$ correspondiente al elemento (i, i); así, en nuestro caso se tiene:

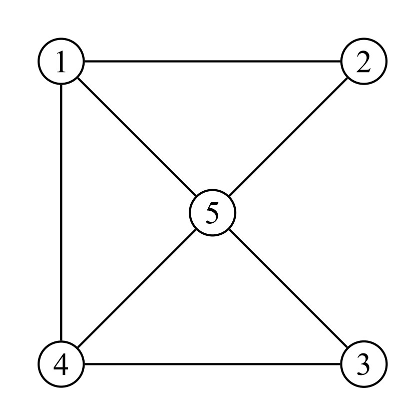
$$Ad(G) = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}, \quad Ad(G - v) = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

Definición 1.2.10
Sea G un grafo simple, y sea $e \in E(G)$. Llamamos **supresión en G de la arista** e al grafo $G - e$ definido como $V(G - e) = V(G)$, $E(G - e) = E(G) - \{e\}$.

Ejemplo 1.2.11
Sea el grafo G definido en el ejemplo anterior, y sea la arista $e = \{2, 3\} \in E(G)$, el grafo $G - e$ tiene como conjunto de vértices $V(G - e) = \{1, 2, 3, 4, 5\}$ y como conjunto de aristas $E(G - e) = \{\{1, 2\}, \{1, 4\}, \{1, 5\}, \{2, 5\}, \{3, 4\}, \{3, 5\}, \{4, 5\}\}$.

En la siguiente representación la figura que aparece a la izquierda representa G, siendo la que aparece a la derecha una representación de $G - e$.



,


♦

La matriz de adyacencia correspondiente a la supresión en G de la arista $e = \{i, j\}$ es la matriz que resulta de anular en $Ad(G)$ los elementos (i, j) y (j, i); así, en nuestro caso se tiene

$$Ad(G) = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}, \quad Ad(G - e) = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & \mathbf{0} & 0 & 1 \\ 0 & \mathbf{0} & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Definición 1.2.11
Sea G un grafo simple, y sea $e = \{i, j\} \notin E(G)$, $i, j \in V(G)$. Llamamos **adición en** G **de la arista** e al grafo $G + e$ definido como suma $G + e = G + G'$ donde G' es el grafo $V(G') = \{i, j\}$, $E(G') = \{\{i, j\}\}$.

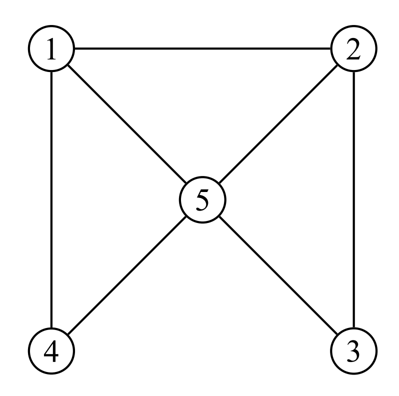
Ejemplo 1.2.12
Sea el grafo G cuyo conjunto de vértices es $V(G) = \{1, 2, 3, 4, 5\}$ y cuyo conjunto de aristas es $E(G) = \{\{1, 2\}, \{1, 4\}, \{1, 5\}, \{2, 5\}, \{3, 5\}, \{4, 5\}\}$, y sea la arista $e = \{3, 4\} \in E(G)$, el grafo $G + e$ tiene como conjunto de vértices:

$$V(G + e) = \{1, 2, 3, 4, 5\}$$

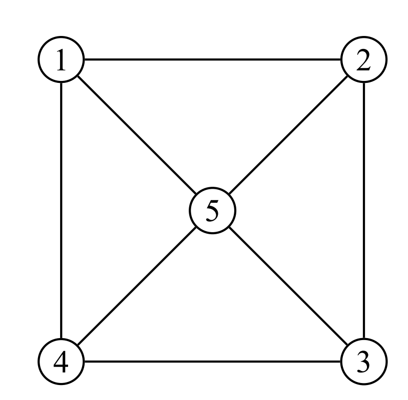
y como conjunto de aristas:

$$E(G + e) = \{\{1, 2\}, \{1, 4\}, \{1, 5\}, \{2, 5\}, \{3, 4\}, \{3, 5\}, \{4, 5\}\}$$

En la siguiente representación la figura que aparece a la izquierda representa G, siendo la que aparece a la derecha una representación de $G + e$.



,



♦

La matriz de adyacencia correspondiente a la adición en G de la arista $e = \{i, j\}$ es la matriz que resulta de hacer $a_{i,j} = a_{j,i} = 1$ en $Ad(G)$. Así, en nuestro caso se tiene

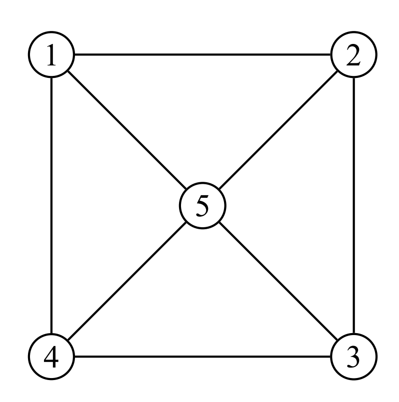
$$Ad(G) \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}, \quad Ad(G + e) = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & \mathbf{1} & 0 & 1 \\ 0 & \mathbf{1} & 0 & 1 & 1 \\ 10 & 1 & 0 & 1 & \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

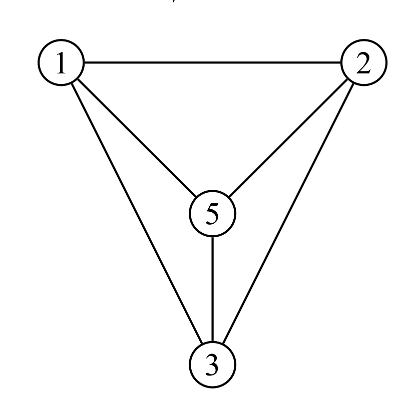
Definición 1.2.12
Sea G un grafo simple y sea $e = \{i, j\} \in E(G)$. Llamamos **contracción del grafo G según la arista** e al grafo G/e definido como $V(G/e) = V(G) - \{v\}$, $E(G/e) = \{\{k, l\} \in E(G) | k \notin \{i, j\}\} \cup \{\{k, j\} \in E(G) | \{k, i\} \in E(G) - \{e\}\}$.

La operación contracción del grafo G según una arista consiste en identificar los vértices i, j extremos de la arista $e = \{i, j\}$, resultando un único vértice. Esta operación implica, por una parte, la eliminación de la arista $e = \{i, j\}$, y por otra, dado que $E(G)$ es un conjunto, la eliminación de elementos repetidos.

Ejemplo 1.2.13
Sea el grafo G definido en el ejemplo anterior, y sea la arista $e = \{3, 4\} \in E(G)$. El grafo G/e tiene como conjunto de vértices $V(G/e) = \{1, 2, 3, 5\}$ y como conjunto de aristas $E(G/e) = \{\{1, 2\}, \{1, 3\}, \{1, 5\}, \{2, 5\}, \{3, 5\}\}$.

En la siguiente representación la figura que aparece a la izquierda representa G, siendo la que aparece a la derecha una representación de G/e.



,


♦

La matriz de adyacencia del grafo G/e, donde $e = \{i, j\}$, $i \leq j$, es la matriz menor complementaria (j, j) de la matriz que resulta de cambiar en $Ad(G)$ la fila (y columna) i por la suma lógica $\vee$ de las filas (columnas) i y j, la cual se efectúa según: $0 \vee 0 = 0$, $0 \vee 1 = 1$, $1 \vee 0 = 1$, $1 \vee 1 = 1$.

La siguiente figura muestra en su parte izquierda la matriz de adyacencia de G y a su derecha la de G/e.

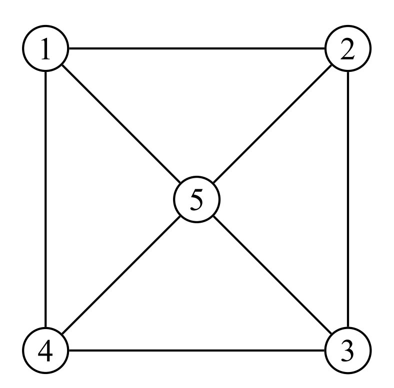
$$Ad(G) = \begin{bmatrix} 0 & 1 & \mathbf{0} & \mathbf{1} & 1 \\ 1 & 0 & \mathbf{1} & \mathbf{0} & 1 \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} \\ 1 & 1 & \mathbf{1} & \mathbf{1} & 0 \end{bmatrix}, \quad Ad(G/e) = \begin{bmatrix} 0 & 1 & \mathbf{1} & 1 \\ 1 & 0 & \mathbf{1} & 1 \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} \\ 1 & 1 & \mathbf{1} & 0 \end{bmatrix}$$

Definición 1.2.13
Sea G un grafo simple. Llamamos **grafo lineal** de G al grafo $L(G)$ definido como $V(L(G)) = E(G)$, $E(L(G)) = \{\{e_1, e_2\} \subset E(G) | e_1 \neq e_2,\ e_1, e_2 \text{ incidentes}\}$.

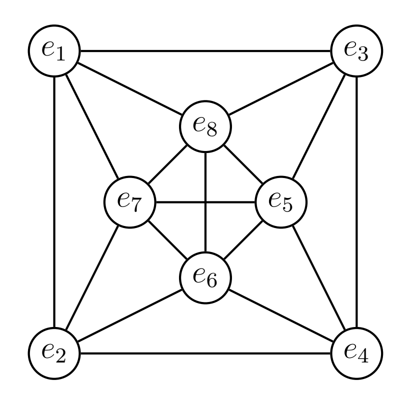
Ejemplo 1.2.14
Consideremos G el grafo de los ejemplos anteriores. El grafo lineal $L(G)$ viene dado por $V(L(G)) = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8\}$, $E(L(G)) = \{\{e_1, e_2\}, \{e_1, e_4\}, \{e_1, e_5\}$

$, \{e_1, e_8\}, \{e_2, e_3\}, \{e_4, e_5\}, \{e_2, e_6\}, \{e_3, e_4\}, \{e_3, e_6\}, \{e_3, e_7\}, \{e_4, e_7\}, \{e_4, e_8\}, \{e_5, e_6\},$
$, \{e_5, e_7\}, \{e_5, e_8\}, \{e_6, e_7\}, \{e_6, e_8\}, \{e_7, e_8\}\}$.

En la siguiente representación la figura que aparece a la izquierda representa G, siendo la que aparece a la derecha una representación de $L(G)$.



,



♦

Para finalizar este apartado, indicar que las operaciones supresión de vértice, supresión de arista, contracción de arista, adición de arista en un grafo G pueden ser fácilmente extensibles a un conjunto de vértices o aristas. Así, sea $\{i_1, \dots, i_k\} \in V(G)$ un conjunto de vértices y $\{e_1, e_2, \dots, e_k\} \in E(G)$ un conjunto de aristas, $\{\{i_1, j_1\}, \dots, \{i_k, j_k\} | \{i_r, j_r\} \notin E(G),\ r = 1, \dots, k\}$, entonces tenemos:

Definición 1.2.14
Con la notación anterior se tiene:

- $G - \{i_1, \dots, i_k\} = (G - \{i_1, \dots, i_{k-1}\}) - i_k$, siendo $G - \{i\} = G - i$.
- $G - \{e_1, \dots, e_k\} = (G - \{e_1, \dots, e_{k-1}\}) - e_k$ siendo $G - \{e\} = G - e$.
- $G/\{e_1, \dots, e_k\} = (G/\{e_1, \dots, e_{k-1}\})/e_k$ siendo $G/\{e\} = G/e$.
- $G + \{\{i_1, j_1\}, \dots, \{i_k, j_k\}\} = (G + \{\{i_1, j_1\}, \dots, \{i_{k-1}, j_{k-1}\}\}) + \{i_k, j_k\}$, siendo $G + \{e\} = G + e$.

1.2.4. Grafos con nombre propio

A continuación se enumeran una serie de grafos de uso común, los cuales tienen nombre propio; así tenemos:

- **Grafo nulo.** Llamamos grafo nulo de k vértices al grafo N_k cuyo conjunto de vértices viene dado por $V(N_k) = \{1, 2, \dots, k\}$ y cuyo conjunto de aristas es vacío, $E(N_k) = \emptyset$. Este grafo carece de aristas, y todos sus vértices son aislados.

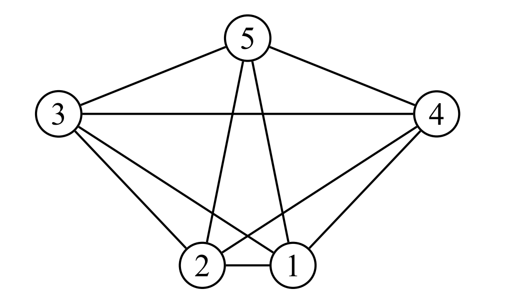
  Ejemplo 1.2.15
  A modo de ejemplo representaremos a continuación el grafo nulo de cuatro vértices N_4.

$$Ad(N_4) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

♦

- **Grafo completo.** Un grafo $G = (V(G), E(G))$ es completo si $(v, w) \in E(G)$ $\forall v, w \in V(G)$. El grafo completo de n vértices se suele denotar por K_n y tiene $\frac{n(n-1)}{2}$ aristas.

  Ejemplo 1.2.16
  A continuación se representa el grafo K_5, y su matriz de adyacencia viene dada por:

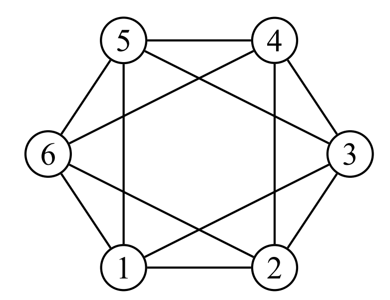


$$Ad(K_5) = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

♦

- **Grafo regular.** Diremos que un grafo G es regular si todos sus vértices tienen el mismo grado.

  Ejemplo 1.2.17
  El grafo que se muestra a continuación es regular, pues todos sus vértices tienen grado cuatro.

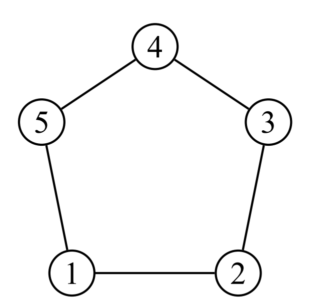


$$Ad(G) = \begin{bmatrix} 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$

♦

- **Grafo circuito.** Grafo conexo donde todos sus vértices tienen grado dos. Se representa por C_n.

  Ejemplo 1.2.18
  A continuación se muestra una representación del grafo circuito de cinco vértices, así como su matriz de adyacencia.

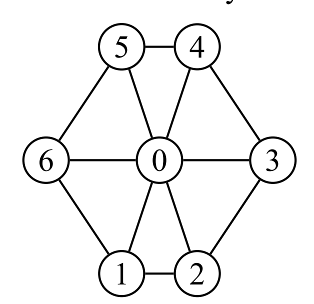


$$Ad(C_5) = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

♦

- **Grafo rueda de n radios.** Es el grafo suma de N_1 y C_n; se representa por R_n.

  Ejemplo 1.2.19
  La figura siguiente muestra el grafo rueda de cinco radios junto a la correspondiente matriz de adyacencia.



$$Ad(R_6) = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

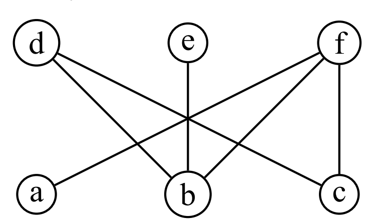
♦

- **Grafo bipartido.** Diremos que un grafo G es un grafo bipartido si existen dos subconjuntos disjuntos no vacíos $V_1 \subset V(G)$, $V_2 \subset V(G)$ de cardinales n_1, n_2 respectivamente, de modo que $V(G) = V_1 \cup V_2$, y $E(G) \subset E(N_{n_1} + N_{n_2})$. Si además se verifica $E(G) = E(N_{n_1} + N_{n_2})$, diremos que el grafo bipartido es completo, en cuyo caso se denotará como K_{n_1,n_2}.

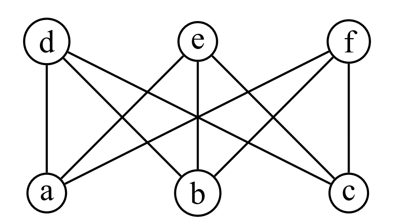
  Nótese que el concepto de grafo bipartido equivale a decir que es posible establecer una partición del conjunto de vértices en dos subconjuntos de modo que todas las aristas del grafo tienen un extremo en el primer subconjunto y otro extremo en el segundo subconjunto.

  Otra definición equivalente de grafo bipartido es la siguiente: subgrafo de $n_1 + n_2$ vértices de $N_{n_1} + N_{n_2}$.

  Ejemplo 1.2.20
  A continuación se muestran dos grafos bipartidos, el primero de ellos incompleto y el segundo completo. A la derecha de ambos grafos aparece su matriz de adyacencia.



$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$



$$Ad(K_{3,3}) = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

♦

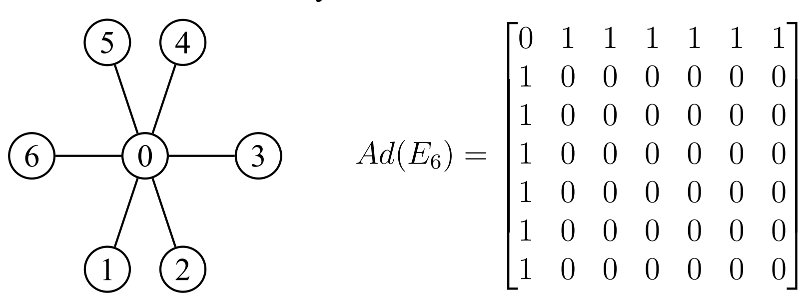
La matriz de adyacencia de un grafo bipartido toma la forma

$$\begin{bmatrix} 0_{n_1} & A \\ A^t & 0_{n_2} \end{bmatrix}$$

donde 0_k representa la matriz nula $k \times k$, y donde A es una matriz $n_1 \times n_2$, siendo A^t su traspuesta. En el caso de ser grafo bipartido completo todos los elementos de la submatriz A serán la unidad.

- **Estrella de n puntas.** Es el grafo suma $N_1 + N_n$, el cual se denota como E_n.

  Ejemplo 1.2.21
  A continuación se muestra como ejemplo una representación del grafo estrella de seis puntas, así como su matriz de adyacencia.



$$Ad(E_6) = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

♦

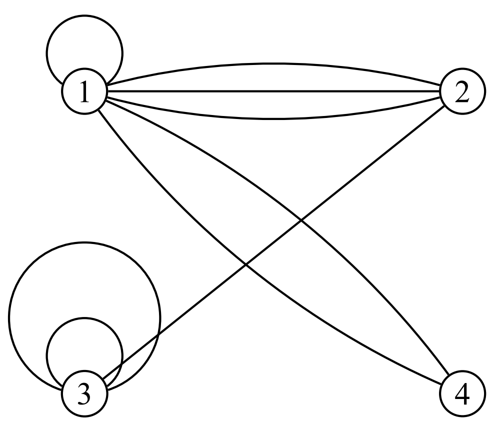
1.2.5. Grafos generales

En la definición de grafo, se definió $E(G)$ utilizando el término conjunto, lo cual impide la posibilidad de tener elementos repetidos. Esto supone la siguiente restricción: en un grafo simple sólo puede haber una arista que una dos vértices. También se definió arista como subconjunto de dos elementos de $V(G)$, lo cual impide aristas del tipo $\{a, a\}$.

El concepto de grafo puede ser extendido de modo que quepan tanto aristas múltiples como aristas de tipo $\{a, a\}$.

Definición 1.2.15
Un **grafo general** G es un par $(V(G), E(G))$ donde $V(G)$ es un conjunto finito no vacío cuyos elementos llamamos vértices y $E(G)$ una familia cuyos elementos son familias de dos elementos de $V(G)$ a las cuales llamamos aristas.

Ejemplo 1.2.22
A continuación se muestra un ejemplo de grafo general.



El grafo G de la figura tiene como conjunto de vértices

$$V(G) = \{1, 2, 3, 4\}$$

y como familia de aristas

$$E(G) = \{\{1,1\}, \{1,2\}, \{1,2\}, \{1,2\}, \{1,4\}, \{1,4\}, \{2,3\}, \{3,3\}, \{3,3\}\}$$

La matriz de adyacencia del grafo viene dada por:

$$Ad(G) = \begin{bmatrix} 2 & 3 & 0 & 2 \\ 3 & 0 & 1 & 0 \\ 0 & 1 & 4 & 0 \\ 2 & 0 & 0 & 0 \end{bmatrix}$$

♦

El concepto de familia de elementos permite la existencia de elementos repetidos, considerándose dos familias iguales si únicamente varía el orden en el que están colocados sus miembros. En caso de arista múltiple las podemos distinguir denotándolas como $\{i, j\}_1$,$\{i, j\}_2$,. . .,$\{i, j\}_k$.

En un grafo general llamaremos lazo a las aristas cuyos extremos coinciden. Para calcular el orden de un vértice cada lazo cuenta como dos.

En un grafo general puede definirse, de modo análogamente al caso de grafo simple, la matriz de adyacencia $Ad(G)$, siendo en este caso $a_{i,j}$ el número de aristas que tienen por extremos los vértices i, j. En este caso debe considerarse cada lazo como dos.

En este curso, cuando nos refiramos a grafo lo haremos a grafo simple, debiendo explicitarse cuando se haga referencia a grafo general.

1.2.6. Dígrafos

Hasta el momento hemos considerado que en un grafo G, las aristas son un subconjunto (o una clase) de dos elementos de $V(G)$, lo cual hace equivalente $\{i, j\}$ a $\{j, i\}$. Si se quiere considerar el sentido en el que pueden ser recorridas las aristas, el modelo hasta ahora considerado no es suficiente, por lo que resulta necesaria la definición de un nuevo concepto, el de grafo dirigido o dígrafo, en el cual se tiene en cuenta esta consideración. Así se tiene:

Definición 1.2.16
Se dice que D es un **dígrafo simple** si D es un par $D = (V(D), A(D))$ donde $V(D) = \{1, 2, \ldots, n\}$ es un conjunto finito no vacío cuyos elementos se llaman vértices (o nodos) y $A(D)$ es un subconjunto de $V(D) \times V(D) - \Delta_{V((D)\times V(D)}$ donde $V(D) \times V(D)$ representa el producto cartesiano de $V(D)$ por sí mismo y $\Delta_{V((D)\times V(D)} = \{(i,i)|i \in V(D)\}$, la diagonal de dicho producto cartesiano. A los elementos de $A(D)$ se les llama arcos (también aristas dirigidas o diaristas).

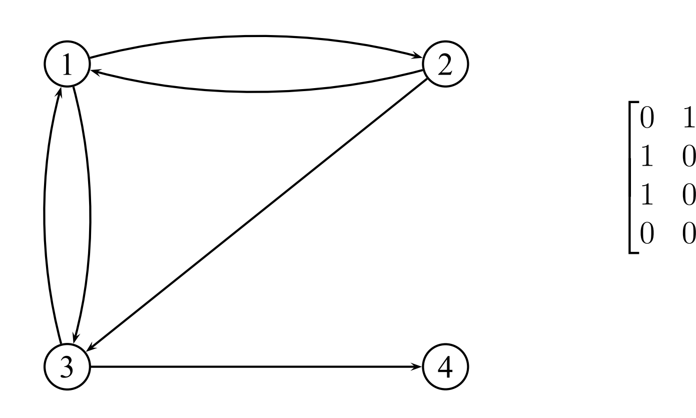
En la representación de un dígrafo se utilizan flechas para indicar el sentido de los arcos. También es común la aparición de segmentos no dirigidos en las representaciones, lo que debe interpretarse en el siguiente sentido: si en un dígrafo aparece una arista $\{i, j\}$ debe entenderse que dicho dígrafo posee como arcos tanto a (i, j) como a (j, i).

En un dígrafo, un vértice i es adyacente a un vértice j si $(i, j) \in A(D)$. En este caso, la adyacencia no tiene la propiedad de simetría. Este hecho se refleja en la matriz de adyacencia $Ad(D)$, la cual es una matriz $n \times n$ donde n es el número de vértices que contiene $V(D)$ y cuyos elementos vienen dados por:

$$Ad(D) = \begin{bmatrix} a_{1,1} & \ldots & a_{i,n} \\ \ldots & \ldots & \ldots \\ a_{n,1} & \ldots & a_{n,n} \end{bmatrix}, \qquad a_{i,j} = \begin{cases} 1 & (i,j) \in A(D) \\ 0 & (i,j) \notin A(D) \end{cases}$$

La matriz de adyacencia de un dígrafo no es necesariamente simétrica.

Ejemplo 1.2.23
A continuación se muestra la representación gráfica de un dígrafo simple y su matriz de adyacencia.



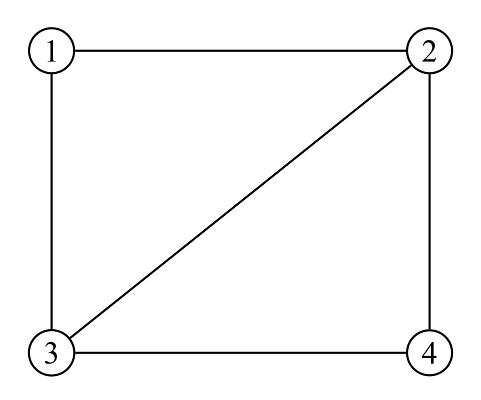
$$\begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

♦

A cada dígrafo D se le puede asociar un grafo G_S llamado **grafo subyacente**, el cual resulta de transformar los arcos en aristas, eliminando en su caso aristas múltiples, y que se define formalmente como:

Definición 1.2.17
Sea D un dígrafo. Llamamos grafo subyacente a D a un grafo G definido como $V(G) = V(D)$, $E(G) = \bigcup_{(i,j)\in A(D)} \{\{i, j\}\}$.

Ejemplo 1.2.24
Sea D el dígrafo del ejemplo anterior, su grafo subyacente G viene representado por:



♦

Por otra parte se tiene que si $Ad(D) = [d_{i,j}]_{i,j=1}^{n}$ es la matriz de adyacencia de D, entonces la matriz de adyacencia del grafo subyacente G viene dada por $Ad(G) = [a_{i,j}]_{i,j=1}^{n}$, donde $a_{i,j} = a_{j,i} = \text{máx}\{d_{i,j}, d_{j,i}\}$.

Al igual que en el caso de grafos, la palabra dígrafo la reservaremos para el caso de un dígrafo simple. En el caso de ser necesaria la inclusión de arcos múltiples o de lazos, el concepto de dígrafo debe ser ampliado. Por ello se introduce el concepto de dígrafo general, el cual puede contener arcos múltiples y lazos, y que se define como:

Definición 1.2.18
Se dice que D es un **dígrafo general** si D es un par $D = (V(D), A(D))$ donde $V(D) = \{1, 2, \ldots, n\}$ es un conjunto finito no vacío cuyos elementos se llaman vértices (o nodos) y $A(D)$ es una familia de elementos de $V(D) \times V(D)$.

1.3. Recorridos en grafos y dígrafos. Conexión

Para estudiar propiedades de los grafos, es interesante el estudio de los distintos modos de ir desde un vértice a otro, para a continuación estudiar propiedades respecto a la conectividad. Para ello, dado un grafo $G = (V(G); E(G))$ se definen los siguientes conceptos:

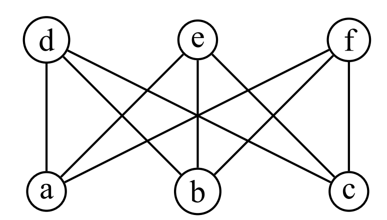
1.3.1. Secuencias de aristas, colas y trayectorias

Definición 1.3.1
Dado un grafo G, llamamos **secuencia de aristas** de G a una familia de vértices $\{i_1, i_2, \ldots, i_k\}$ de modo que $\{i_j, i_{j+1}\} \in E(G)$. Dicha secuencia se dice que conecta los vértices i_1 e i_k y se representa por $i_1 \to i_2 \to \cdots \to i_k$. En el caso de que $i_1 = i_k$, la secuencia de aristas se dice cerrada.

La secuencia de aristas está formada por las aristas $\{i_1, i_2\}$, $\{i_2, i_3\}$,. . .,$\{i_{k-1}, i_k\}$. En una secuencia de aristas los vértices e_j, $e_{j+1}\,\forall j = 1, \ldots, k$ son adyacentes. Finalmente, indicar que en una secuencia de aristas pueden repetirse vértices y aristas.

La definición anterior puede extenderse a dígrafos sin más que cambiar aristas por arcos. En este caso se llama secuencia de diaristas o secuencia de arcos.

Ejemplo 1.3.1
Sea el grafo G



una secuencia de aristas en G puede ser

$$a \to e \to a \to f \to c \to e \to a \to d$$

♦

Definición 1.3.2
Sea G un grafo. Llamamos **cola** a una secuencia de aristas en la que todas las aristas son diferentes. Cuando el primer y último vértice de la cola coinciden, se dice que la cola es **cerrada**.

En una cola pueden repetirse vértices pero no aristas.

La definición anterior también puede extenderse de modo natural a dígrafos cambiando aristas por arcos. En este caso se llama cola dirigida o dicola.

Ejemplo 1.3.2
En el grafo anterior una posible cola puede ser

$$a \to e \to b \to d \to a \to f \to c \to d$$

♦

Definición 1.3.3
Sea G un grafo. Una **trayectoria** es una cola en la que además todos los vértices son diferentes, excepto a lo sumo el primero y el último.

En una trayectoria no puede haber vértices iguales, excepto a lo sumo el primero y el último, ni tampoco aristas repetidas.

La definición anterior puede extenderse a dígrafos sin más que cambiar aristas por arcos. En este caso se llama trayectoria dirigida o ditrayectoria.

Ejemplo 1.3.3

$$a \to e \to b \to f \to c \to d$$

♦

Definición 1.3.4
Sea G un grafo. Un **circuito** en G es una trayectoria cerrada.

En el caso de dígrafos se puede extender la definición anterior sin más que cambiar aristas por arcos. En este caso recibe el nombre de circuito dirigido o dicircuito.

Ejemplo 1.3.4

$$a \to e \to b \to f \to c \to d \to a$$

♦

Definición 1.3.5
Se llama **número de pasos** de una secuencia de aristas (cola, trayectoria, circuito, disecuencia...) al número de aristas (arcos) que la compone.

Consideremos un grafo G cuya matriz de adyacencia $Ad(G)$ representaremos por A. La matriz de adyacencia indica el número de secuencias de aristas de longitud uno que conectan el vértice i con el vértice j, pues $a_{i,j}$ toma valor cero o uno según exista la arista $\{i, j\}$ o no. Este resultado se puede extender del siguiente modo:

Teorema 1.3.1
Sea A la matriz de adyacencia de un grafo simple G. Entonces A^k tiene por elemento (i, j) el número de secuencias de aristas de k pasos distintas que conectan los vértices i y j.

Dem:

Para $n = 0$ se tiene que $A^0 = I$ es la matriz identidad, por lo que la afirmación se cumple. Para $n = 1$ se cumple como hemos visto anteriormente.

Supongamos que se cumple para un k. Entonces el número de secuencias de aristas de longitud $k + 1$ distintos que conectan i y j lo podemos determinar como la suma del número de secuencias de longitud k que conectan i con los vértices adyacentes a j.

Puesto que la matriz A es simétrica, se verifica que A^k también lo es. Así se tiene que el elemento (i, j) de A^{k+1}, que denotaremos por $a_{i,j}^{(k+1)}$, estará determinado como:

$$a_{i,j}^{(k+1)} = \sum_{r=1}^{n} a_{i,r}^{(k)} a_{r,j} = \sum_{r \in V_j} a_{i,r}^{(k)}$$

donde $n = \text{card}\,(V(G))$ y $V_j = \{s \in V(G) | a_{s,j} = 1\}$ el conjunto de vértices adyacentes a j, con lo que se tiene resultado. ▼

1.3.2. Conexión

Definición 1.3.6
Sea G un grafo simple. Diremos que G es **conexo** si para cualquier par de vértices i, j de $V(G)$ existe una trayectoria que tiene por vértice inicial a i y por vértice final a j.

Definición 1.3.7
Sea G un grafo simple, y sea $i \in V(G)$. Llamamos **componente conexa** de G que contiene a i, al mayor subgrafo conexo de G que contiene a i. La componente conexa de G que contiene a i se denota por G_i;

$$G_i = \{j \in V(G) | \exists i_1, \ldots, i_{s-1} \in V(G); \{i_{k-1}, i_k\} \in E(G),\ k = 0, \ldots, s\},$$

siendo $i_0 = i$, $i_s = j$.

Todo grafo no conexo G puede obtenerse como unión de sus componentes conexas. Para ello podemos proceder como sigue:

1. Tomar el conjunto de aristas $E(G)$, de modo que si $\{1, j\} \in E(G)$ se mantenga $i < j$, y asignar inicialmente a cada vértice $i \in V(G)$ del grafo, un entero $c(i) = i$, de modo que en cada momento $(c(1), c(2), \ldots, c(n))$ determina a qué componente conexa pertenece cada vértice.

2. Recorrer el conjunto de aristas tomando una arista $\{i, j\}$ en cada paso.

3. Si $c(i) = c(j)$ ir al paso siguiente. Si $c(i) \neq c(j)$ hacer para $k = 1, \ldots, n$

$$c(k) = \begin{cases} c(k) & \text{si } c(k) < c(j) \\ c(i) & \text{si } c(k) = c(j) \\ c(k) - 1 & \text{si } c(k) > c(j) \end{cases}.$$

4. Si el conjunto de aristas no está totalmente recorrido ir al paso 2.

5. En este punto finaliza el algoritmo resultando el número de componentes conexas $nc = \max\{c(i) | i = 1, \ldots n\}$. La componente conexa G_k, $k = 1, \ldots, nc$ tiene por vértices $V(G_k) = \{i \in V(G) | c(i) = k\}$ y por aristas $E(G_k) = \{\{i, j\} \in E(G) | n(i) = n(j) = k\}$.

Ejemplo 1.3.5
Sea el grafo G cuya representación y matriz de adyacencia $A = Ad(G)$ vienen dadas por

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Las aristas del grafo G viene dadas por $E(G) = \{\{1, 2\}, \{1, 3\}, \{2, 3\}, \{4, 5\}\}$. La aplicación del algoritmo anterior conduce a la tabla

arista	$c(1)$	$c(2)$	$c(3)$	$c(4)$	$c(5)$
	1	2	3	4	5
{1,2}	1	1	2	3	4
{1,3}	1	1	1	2	3
{2,3}	1	1	1	2	3
{4,5}	1	1	1	2	2

Por tanto, el grafo G, tiene dos componentes conexas G_1, G_2 definidas por $V(G_1) = \{1, 2, 3\}$, $E(G_1) = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$ y $V(G_2) = \{4, 5\}$, $E(G_2) = \{\{4, 5\}\}$. ♦

La conexión también puede ser caracterizada mediante la matriz de adyacencia a partir del siguiente teorema:

Teorema 1.3.2
Sea G un grafo simple de n vértices, cuya matriz de adyacencia es A. Entonces, G es conexo si, y sólo si, la matriz $\sum_{k=0}^{n-1} A^k$ tiene todos sus elementos distintos de cero.

Dem:

- $\Rightarrow$) Si un grafo es conexo, dados dos vértices distintos cualesquiera v_i, v_j existe al menos una trayectoria de menos de n aristas que conecta ambos vértices. Sea $0 < k \leq n-1$ la longitud de una de tales trayectorias, entonces $a_{i,j}^{(k)} \geq 1$, por tanto $\sum_{r=0}^{n-1} a_{i,j}^{(r)} \geq a_{i,j}^{(k)} \geq 1 > 0$. De aquí, $\sum_{r=0}^{n-1} A^r$ tiene todos los coeficientes no nulos.

- $\Leftarrow$) Si $\sum_{r=0}^{n-1} A^r$ tiene todos sus elementos no nulos, se tiene que $\forall i, j \in \{1, \ldots, n\}$, $i \neq j$, $\sum_{r=0}^{(k)} > 0$ y puesto que $a_{i,j}^{(0)} = 0$, si $i \neq j$ y $a_{i,j}^{(r)} \geq 0\ \forall r \in \{1, \ldots, n-1\}$ se tiene que $\exists k \in \{1, \ldots, n-1\}$ de modo que $a_{i,j}^{(k)} \geq 1$. Por tanto, G es conexo.

▼

Ejemplo 1.3.6
Sea el grafo G dado en el ejemplo anterior. Como se ve en su representación, dicho grafo no es conexo. Para caracterizarlo a partir de la matriz de adyacencia se tiene

que

$$A^0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \qquad A^2 = \begin{bmatrix} 2 & 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 1 & 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A^3 = \begin{bmatrix} 2 & 3 & 3 & 0 & 0 \\ 3 & 2 & 3 & 0 & 0 \\ 3 & 3 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \qquad A^4 = \begin{bmatrix} 6 & 5 & 5 & 0 & 0 \\ 5 & 6 & 5 & 0 & 0 \\ 5 & 5 & 6 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

y por tanto

$$S = \sum_{k=0}^{5-1} A^k = \begin{bmatrix} 11 & 10 & 10 & 0 & 0 \\ 10 & 11 & 10 & 0 & 0 \\ 10 & 10 & 11 & 0 & 0 \\ 0 & 0 & 0 & 3 & 2 \\ 0 & 0 & 0 & 2 & 3 \end{bmatrix}$$

La matriz S tiene elementos nulos (i.e. $s_{1,4} = 0$), por tanto G es no conexo. ♦

Ejemplo 1.3.7
El siguiente grafo G' cuya representación y matriz de adyacencia D aparecen a continuación es conexo;

4 - 3
1 - 2

$$D = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

de donde

$$A^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad A^2 = \begin{bmatrix} 0 & 2 & 2 & 2 \\ 2 & 0 & 2 & 2 \\ 2 & 2 & 0 & 2 \\ 2 & 2 & 2 & 0 \end{bmatrix}, \quad A^3 = \begin{bmatrix} 3 & 4 & 0 & 4 \\ 4 & 3 & 4 & 0 \\ 0 & 4 & 3 & 4 \\ 4 & 0 & 4 & 3 \end{bmatrix}$$

a partir de lo cual se tiene

$$H = \sum_{k=0}^{4-1} D^k = \begin{bmatrix} 4 & 3 & 7 & 3 \\ 3 & 4 & 3 & 7 \\ 7 & 6 & 4 & 3 \\ 3 & 7 & 3 & 4 \end{bmatrix}$$

En este caso, la matriz H tiene todos sus elementos no nulos, lo que implica que el grafo es conexo. ♦

Teorema 1.3.3
Sea G un grafo simple, y sea $A = Ad(G)$ su matriz de adyacencia. Entonces una arista $e = \{i, j\} \in E(G)$ pertenece a un circuito si, y sólo si, se verifica que $s_{i,j} \neq 0$,

siendo S la matriz $S = \sum_{k=0}^{n-1} B^k$ donde B es la matriz de adyacencia del grafo $G - e$ y n es el número de vértices del grafo G.

Dem:

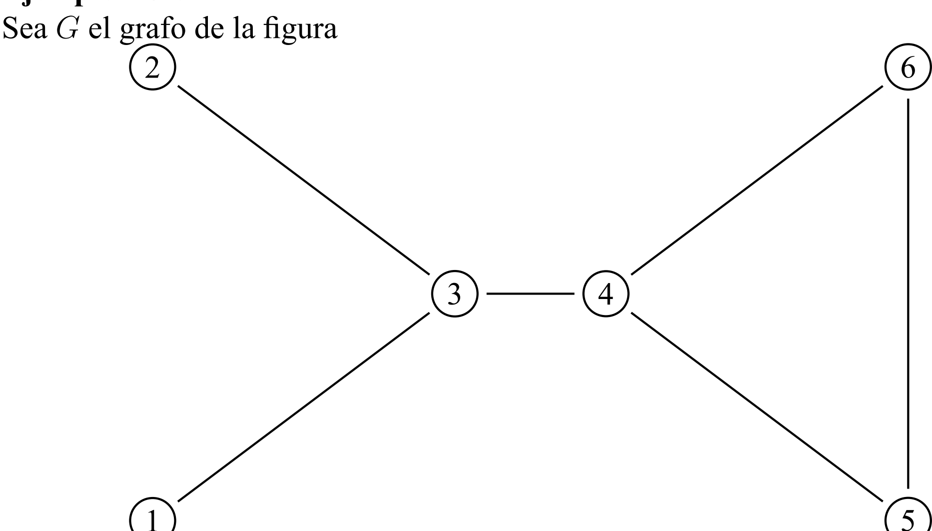
Inmediata, pues para que una arista pertenezca a un circuito es necesario y suficiente que exista un camino del vértice i al vértice j que no contenga a dicha arista, lo que es equivalente a poder conectar los vértices i, j en el grafo $G - e$, lo que a su vez equivale a que el elemento $s_{i,j}$ de la matriz $S = \sum_{k=0}^{n-1} B^k$ sea distinto de cero, siendo B la matriz de adyacencia del grafo $G - e$ y n, el número de vértices del grafo G. ▼

Definición 1.3.8

Sea G un grafo conexo. Se dice que $D \subset E(G)$ es un **conjunto desconectador** de G si el grafo $G = (V(G), E(G) - D)$ es no conexo.

Ejemplo 1.3.8

Sea G el grafo de la figura



y sean los conjuntos $D_1 = \{\{4,5\},\{4,6\}\}$, $D_2 = \{\{5,6\}\}$. El primer conjunto es conjunto desconectador pues $G - D_1$ tiene por matriz de adyacencia

$$B_1 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

a partir de la cual se puede determinar la matriz

$$\sum_{k=0}^{n-1} B_1^k = \begin{bmatrix} 5 & 4 & 13 & 4 & 0 & 0 \\ 4 & 5 & 13 & 4 & 0 & 0 \\ 13 & 13 & 13 & 13 & 0 & 0 \\ 4 & 4 & 13 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 0 & 3 & 3 \end{bmatrix}$$

la cual tiene elementos nulos, siendo por tanto $G - D_1$ no conexo.

El segundo conjunto no es desconectador pues $G - D_2$ tiene por matriz de adyacencia

$$B_2 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

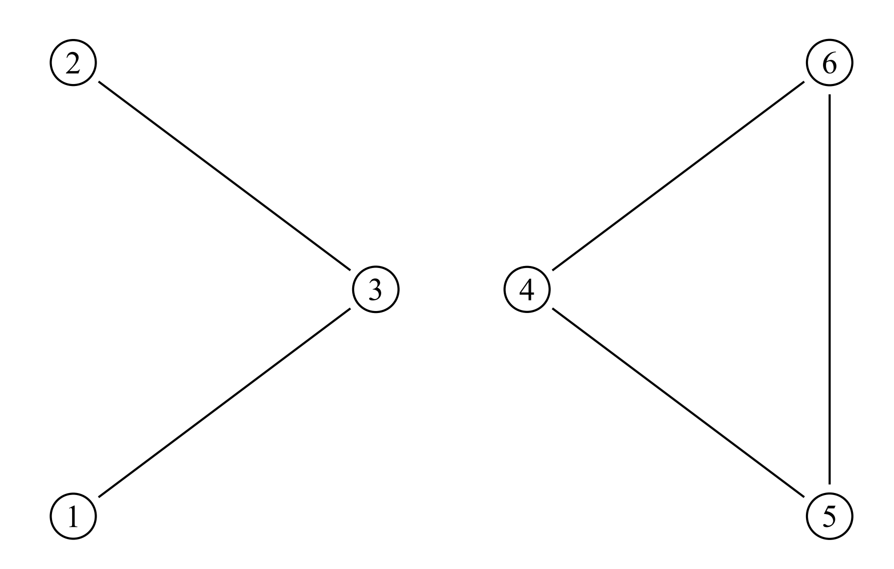
a partir de la cual se puede determinar la matriz

$$\sum_{k=0}^{n-1} B_2^k = \begin{bmatrix} 5 & 4 & 15 & 6 & 6 & 6 \\ 4 & 5 & 15 & 6 & 6 & 6 \\ 15 & 15 & 15 & 27 & 6 & 6 \\ 6 & 6 & 27 & 15 & 15 & 15 \\ 6 & 6 & 6 & 15 & 5 & 4 \\ 6 & 6 & 6 & 15 & 4 & 5 \end{bmatrix}$$

la cual tiene todos sus elementos no nulos, siendo por tanto $G - D_2$ conexo. ♦

Definición 1.3.9
Sea G un grafo conexo, y sea $e \in E(G)$. Diremos que e es un **istmo o puente** si $D = \{e\}$ es conjunto desconectador de G.

Ejemplo 1.3.9
Sea el grafo dado en la figura, la arista $\{3, 4\}$ es un istmo o puente pues su supresión, tal y como se aprecia en la figura, desconecta el grafo



♦

Definición 1.3.10
Sea G un grafo conexo. Diremos que un conjunto de aristas $C \subset E(G)$ es un **conjunto de corte** si es un conjunto desconectador minimal, esto es, C conjunto desconectador de manera que si $K \subset C$ y $K \neq C$ entonces K no es conjunto desconectador.

Ejemplo 1.3.10
Sea el grafo dado en el ejemplo 1.3.8, y sean los conjuntos $D_1 = \{\{4,5\},\{4,6\}\}$ y $D_2 = \{\{1,2\},\{3,4\}\}$.

El primero de ellos es conjunto de corte, pues ningún subconjunto propio suyo lo desconecta G, pues para sus subconjuntos propios $\{\{4,5\}\}$ y $\{\{4,6\}\}$ se tiene que $\{\{4,5\}\}$ no desconecta G, y $\{\{4,6\}\}$ tampoco lo hace. El conjunto D_2 no es conjunto de corte, pues $\{\{3,4\}\}$ desconecta G y es subconjunto propio de D_2.

Definición 1.3.11
Sea G un grafo conexo. Se define **orden de aristo-conectividad** al menor cardinal de sus conjuntos de corte. Esta cantidad se representa por $\lambda(G)$:

$$\lambda(G) = \text{mín}\{\text{card}\,(C)|\ C \text{ conjunto de corte de } G\}.$$

Ejemplo 1.3.11
En el grafo anterior los posibles conjuntos de corte son $C_1 = \{\{1,3\}\}$, $C_2 = \{\{2,3\}\}$, $C_3 = \{\{3,4\}\}$, $C_4 = \{\{4,5\},\{4,6\}\}$, $C_5 = \{\{4,5\},\{5,6\}\}$, $C_6 = \{\{4,6\},\{5,6\}\}$, así

$$\lambda(G) = \text{mín}\{\text{card}\,(C)|\ C \text{ conjunto de corte de } G\} = \text{mín}\{1,1,1,2,2,2\} = 1$$

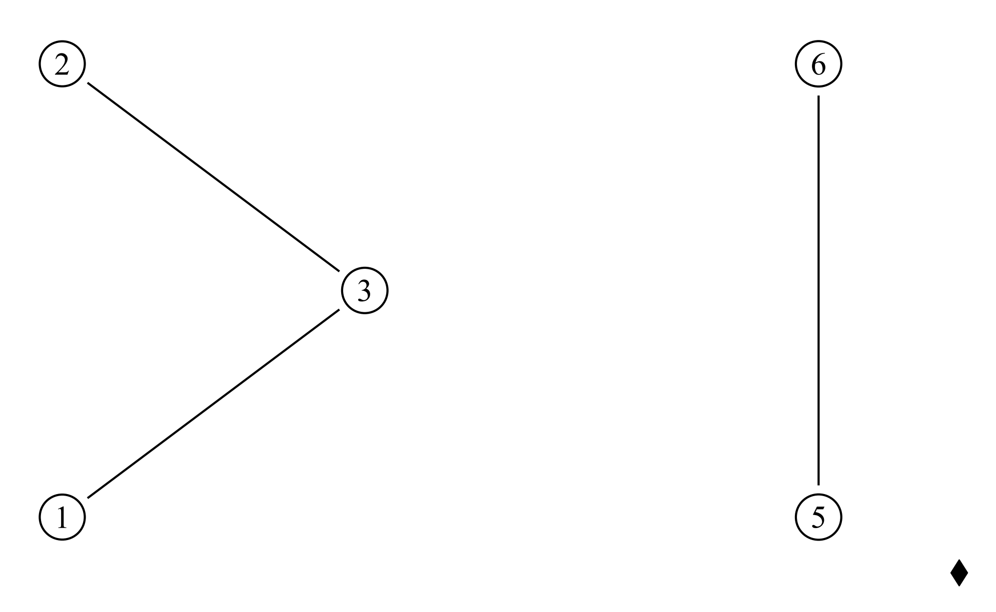
♦

Definición 1.3.12
Sea G un grafo conexo. Un **conjunto separador** de un grafo conexo es un conjunto de vértices $S \subset V(G)$ de modo que el grafo $G - S$ es no conexo.

Ejemplo 1.3.12
En el grafo dado en 1.3.8 los conjuntos $S_1 = \{2,3\}$, $S_2 = \{4\}$ son conjuntos separadores, pues en ambos casos es fácil comprobar que los grafos $G - S_1$ y $G - S_2$ son no conexos. El conjunto $S_3 = \{4,5\}$ no es separador, pues $G - S_3$ resulta conexo. ♦

Definición 1.3.13
Sea G un grafo conexo. Diremos que un vértice $i \in V(G)$ es **vértice de corte** o **vértice articulación** si $\{i\}$ es conjunto separador de G.

Ejemplo 1.3.13
En el grafo dado en 1.3.8 el vértice 4 es un vértice articulación, pues el grafo $G - 4$ es, tal como aparece en la figura, no conexo.



♦

Definición 1.3.14
Sea G un grafo conexo. Se define **orden de vértice-conectividad** al menor cardinal posible para sus conjuntos separadores, lo que se representa por $\kappa(G)$:

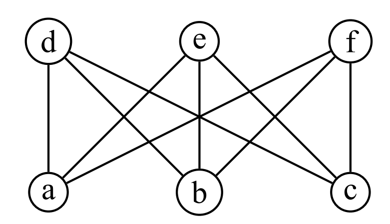
$$\kappa(G) = \text{mín}\{\text{card}\,(S)|S \text{ conjunto separador de } G\}$$

Ejemplo 1.3.14
En el grafo citado en el ejemplo anterior los conjuntos separadores de menor cardinal son $\{3\}$, $\{4\}$. Por tanto, $\kappa(G) = 1$. ♦

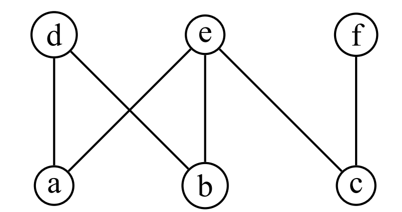
Los conceptos grado de aristo-conectividad y de vértice-conectividad constituyen una medida de la fortaleza de la conexión pues representan el número mínimo de aristas y vértices que pueden ser eliminado (por fallos, destrucción, etc.) para que el sistema resultante falle (deje de ser conexo).

Definición 1.3.15
Sea G un grafo. Diremos que un subconjunto de aristas $C \in E(G)$ es un **conjunto conectador** de G si el subgrafo $(V(G), C)$ es conexo.

Ejemplo 1.3.15
Sea el grafo cuya representación aparece en la figura



El conjunto $\{\{a, d\}, \{a, e\}, \{b, d\}, \{b, e\}, \{c, e\}, \{c, f\}\}$ es conjunto conectador de G pues el grafo $(V(G), S)$ es, tal como aparece a continuación, conexo.

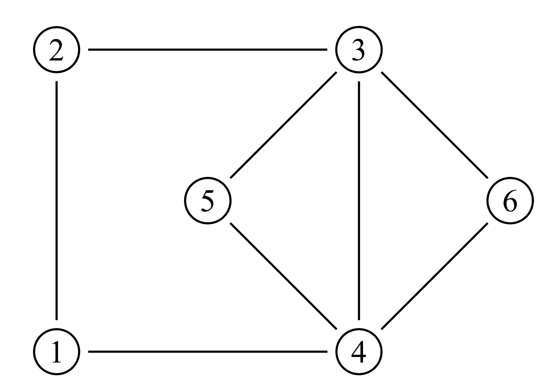


♦

1.4. Grafos eulerianos y hamiltonianos

Definición 1.4.1
Sea G un grafo conexo. Diremos que G es **euleriano** si existe una cola cerrada que incluye todas sus aristas. A dicha cola se le denomina, caso de existir, **cola euleriana**.

Ejemplo 1.4.1
Sea el grafo de la figura

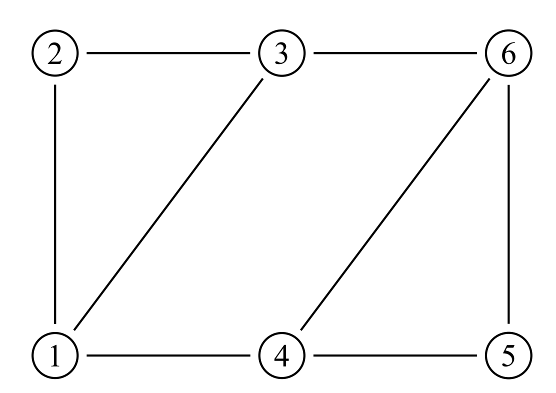


Dicho grafo es euleriano pues admite la cola

$$1 \to 2 \to 3 \to 4 \to 5 \to 3 \to 6 \to 4 \to 1$$

la cual es una cola euleriana.

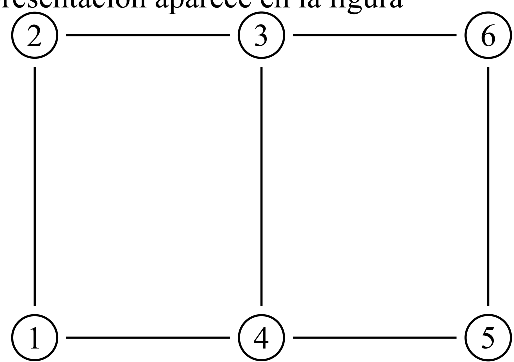
Un ejemplo de grafo no euleriano es el siguiente:



♦

Definición 1.4.2
Sea G un grafo conexo. Diremos que G es **semi-euleriano** si existe una cola que incluye todas sus aristas.

Ejemplo 1.4.2
Sea el grafo cuya representación aparece en la figura



Dicho grafo es semi-euleriano, pues admite la cola

$$3 \to 2 \to 1 \to 4 \to 3 \to 6 \to 5 \to 4$$

♦

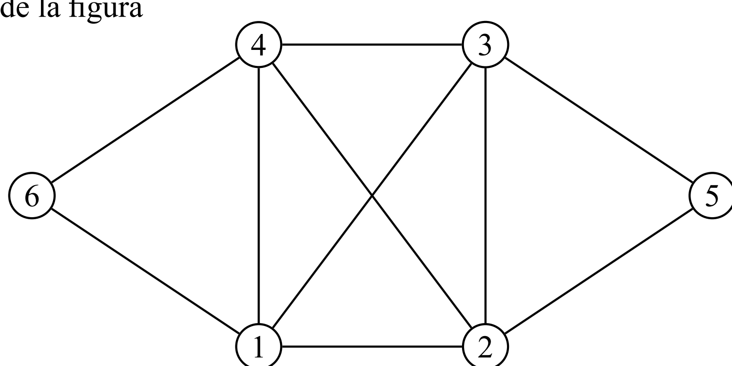
Notar que todo grafo euleriano es semi-euleriano, pero no todo grafo semi-euleriano es euleriano.

Un método para construir una cola euleriana lo proporciona el llamado **algoritmo de Fleury**, el cual puede esquematizarse como sigue:

Elegir un vértice cualquiera v y construir una cola teniendo en cuenta las siguientes reglas:

1. Las aristas se eliminan una vez recorridas.
2. Si un vértice queda aislado, dicho vértice se elimina.
3. No utilizar istmos salvo que no exista otra alternativa.

Ejemplo 1.4.3
Considérese el grafo de la figura



Para tratar de encontrar una cola euleriana en dicho grafo aplicamos el algoritmo de Fleury. Partiendo del vértice 6 se tiene

$$6 \to 1 \to 2 \to 5 \to 3 \to 1 \to 4 \to 2 \to 3 \to 4 \to 6$$

y puesto que existe dicha cola euleriana, puede afirmarse que el grafo es euleriano.
♦

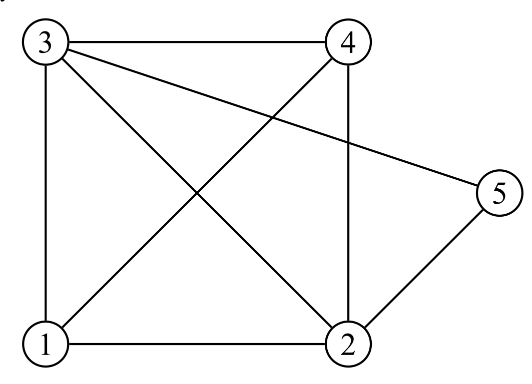
Una caracterización de grafos eulerianos y semi-eulerianos la proporciona los teoremas enunciados a continuación.

Teorema 1.4.1
Un grafo conexo G es euleriano si, y sólo si, el grado de todos sus vértices es par.

Teorema 1.4.2
Un grafo conexo G es semi-euleriano si, y sólo si, el grado de sus vértices, excepto a lo sumo exactamente dos, es par.

En este caso, si existen dos vértices de grado impar el algoritmo de Fleury también proporciona una cola semi-euleriana, debiéndose escoger como vértice inicial uno con grado impar.

Ejemplo 1.4.4
Considérese el grafo de la figura

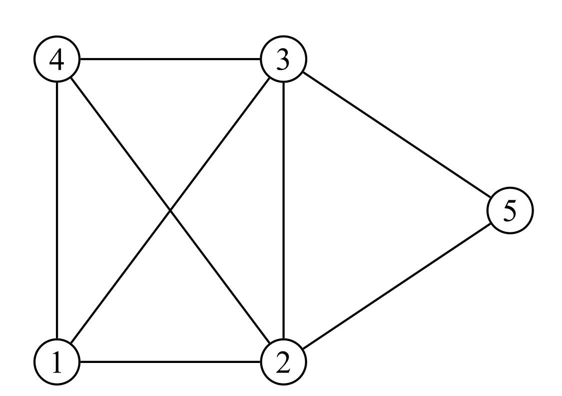


Dicho grafo tiene todos los vértices de grado 4 excepto 2 y 3, que tienen grado 3. Según la caracterización anterior el grafo es semi-euleriano. Una cola semi-euleriana puede ser construida a partir del vértice de orden impar 2. Así se tiene la cola

$$1 \to 2 \to 5 \to 3 \to 1 \to 4 \to 2 \to 3 \to 4$$

♦

Ejemplo 1.4.5
Otro ejemplo de grafo semi-euleriano es el siguiente:

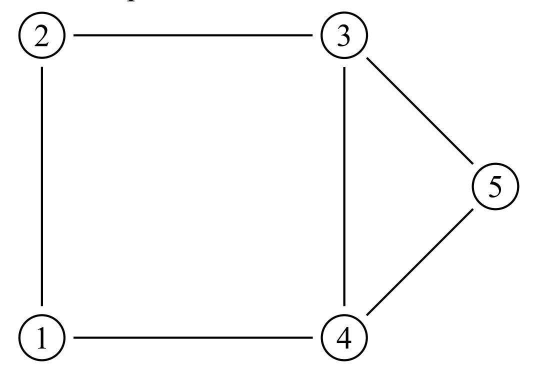


ya que existe la cola $1 \to 4 \to 2 \to 1 \to 3 \to 2 \to 5 \to 3 \to 4$. ♦

A continuación se introduce el concepto de grafo hamiltoniano como:

Definición 1.4.3
Un grafo conexo G se dice **hamiltoniano** si existe un circuito que contenga todos sus vértices.

Ejemplo 1.4.6
El grafo cuya representación aparece a continuación es hamiltoniano

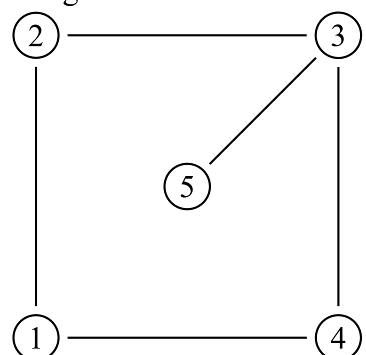


pues admite el circuito

$$1 \to 2 \to 3 \to 5 \to 4 \to 1$$

el cual contiene todos sus vértices.

El grafo que aparece en la figura no es hamiltoniano



pues ningún circuito puede contener el vértice 5. ♦

Definición 1.4.4
Un grafo conexo G se dice semi-hamiltoniano si existe una trayectoria que contenga todos sus vértices.

Ejemplo 1.4.7
El grafo anterior es semi-hamiltoniano, pues la trayectoria

$$2 \to 1 \to 4 \to 3 \to 5$$

contiene todos sus vértices. ♦

A continuación se enuncia sin demostrar una condición suficiente de grafo hamiltoniano:

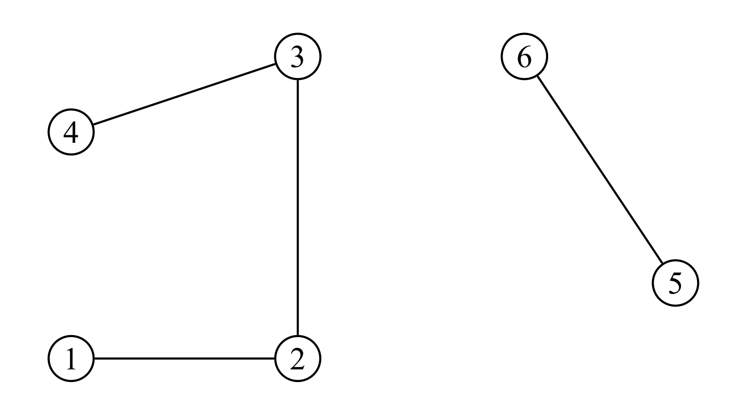
Teorema 1.4.3
Sea G un grafo conexo con n vértices de modo que para cada pareja de vértices no adyacentes v, w se satisfaga $\rho(v) + \rho(w) \geq n$. Entonces G es hamiltoniano.

Lamentablemente, en el caso de grafos hamiltonianos no se dispone de ninguna caracterización (condición necesaria y suficiente).

1.5. Árboles

Definición 1.5.1
Sea G un grafo. Diremos que G es un **bosque** si no contiene ningún circuito.

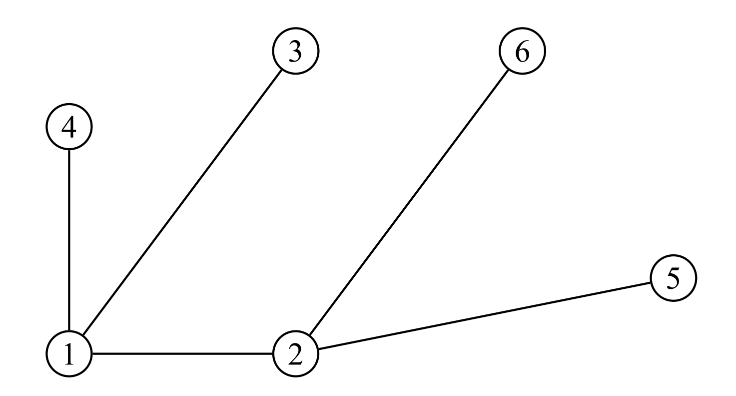
Ejemplo 1.5.1
El grafo que se muestra a continuación es un bosque:



♦

Definición 1.5.2
Diremos que un grafo T es un **árbol** si T es bosque conexo.

Ejemplo 1.5.2
El grafo que se muestra a continuación es un árbol:



♦

Para finalizar, se enuncia el siguiente teorema de caracterización:

Teorema 1.5.1
Sea T un grafo con n vértices. Entonces son equivalentes:

1. T es un árbol.
2. T no contiene ningún circuito y posee $n-1$ aristas.
3. T es conexo y tiene $n-1$ aristas.
4. T es conexo y cada arista es un istmo.
5. Cada dos vértices de T están conectados por una única trayectoria.
6. T no contiene ningún circuito pero la adición de cualquier arista crea exactamente un circuito.

Tema 2

Grafos ponderados y redes

2.1. Introducción

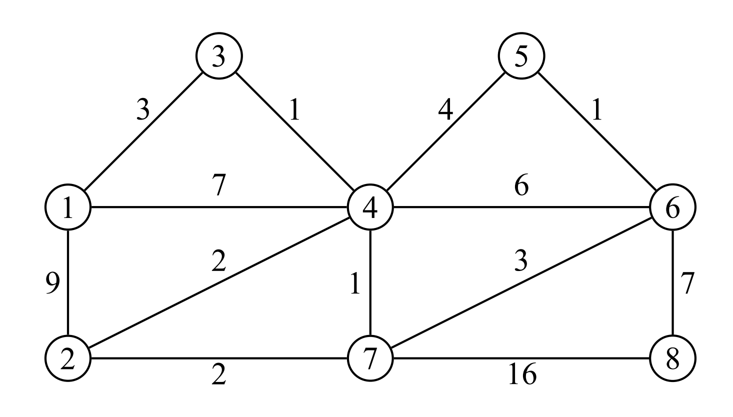
En este capítulo se aborda el estudio de diversos problemas de optimización sobre grafos y dígrafos. En estos problemas se procede a asignar un valor a cada arista/arco de un grafo/dígrafo con el fin de minimizar alguna magnitud calculada sobre el grafo.

Los problemas estudiados en este capítulo serán, el problema de la trayectoria más corta, el problema del conector mínimo, el problema de la trayectoria crítica, y el problema del flujo máximo sobre una red. Para plantear y resolver dichos problemas se proporcionarán las definiciones de los mismos, así como un algoritmo de resolución para cada uno de ellos.

2.2. Grafos ponderados

Definición 2.2.1
Sea G un grafo. Una **ponderación** en G es una aplicación $\omega : E(G) \to \mathbb{R}^+ \cup \{0\}$. En ese caso, el par (G, ω) recibe el nombre de **grafo ponderado** y el valor $\omega(e)$, donde $e \in E(G)$, el de **peso** o **costo** de la arista e.

Ejemplo 2.2.1
El grafo G cuya representación aparece a continuación constituye un ejemplo de grafo ponderado. Los números que aparecen junto a las aristas corresponden a sus pesos.



♦

En teoría de grafos ponderados es conveniente introducir la matriz de pesos como una matriz $\Omega(G)$, $n \times n$ donde $n = card(G)$ de modo que

$$\Omega(G) = \begin{bmatrix} 0 & \omega_{1,2} & \dots & \omega_{1,n-1} & \omega_{1,n} \\ \omega_{2,1} & 0 & \dots & \omega_{2,n-1} & \omega_{2,n} \\ \dots & \dots & \dots & \dots & \dots \\ \omega_{n-1,1} & \omega_{n-1,2} & \dots & 0 & \omega_{n-1,n} \\ \omega_{n,1} & \omega_{n,2} & \dots & \omega_{n,n-1} & 0 \end{bmatrix}$$

$$\text{donde} \quad \omega_{i,j} = \begin{cases} \omega(\{i,j\}) & i \neq j, \quad \{i,j\} \in E(G) \\ +\infty & i \neq j, \quad \{i,j\} \notin E(G) \end{cases}$$

En esta matriz los elementos de la diagonal son cero. En los lugares donde aparece $+\infty$ debe entenderse que no hay arista.

2.2.1. El problema del conector mínimo

El problema del conector mínimo podemos enunciarlo como sigue:

Definición 2.2.2
Dado un grafo G conexo, llamamos **problema del conector mínimo** a encontrar un subgrafo de G de la forma $(V(G), C)$ donde C es un conjunto conectador de G de peso mínimo.

El problema del conector mínimo puede resolverse mediante el llamado algoritmo profundo o de Kruskal, el cual es expuesto a continuación.

Algoritmo de Kruskal:

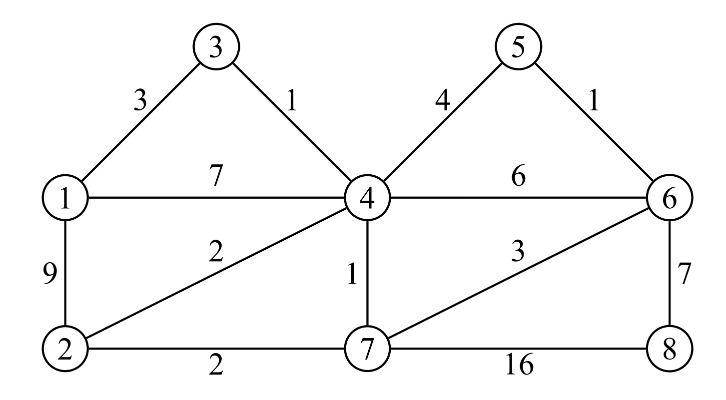
1. Se toma una arista e de peso mínimo, y se definen los conjuntos $C = \{e\}$, $D = \emptyset$.

2. Se toma una arista f en $E(G) - C \cup D$ de peso mínimo.

3. Si $\begin{cases} (V(G), C \cup \{f\}) \text{ contiene un circuito, entonces } D = D \cup \{f\}. \\ (V(G), C \cup \{f\}) \text{ no contiene ningún circuito, entonces } C = C \cup \{f\}. \end{cases}$

4. Si $\begin{cases} \text{card}\,(C) = \text{card}\,(V(G)) - 1 & \text{FIN: el conector mínimo es } (V(G), C). \\ \text{card}\,(D) \neq \text{card}\,(V(G)) - 1 & \text{ir al paso 2.} \end{cases}$

Ejemplo 2.2.2
Dado el grafo G representado en la figura, obtener un conector mínimo.



El conector mínimo se determina mediante el algoritmo Kruskal, para ello, en primer lugar se ordenan las aristas en pesos crecientes, los cuales se indican como subíndice. Así se tiene:

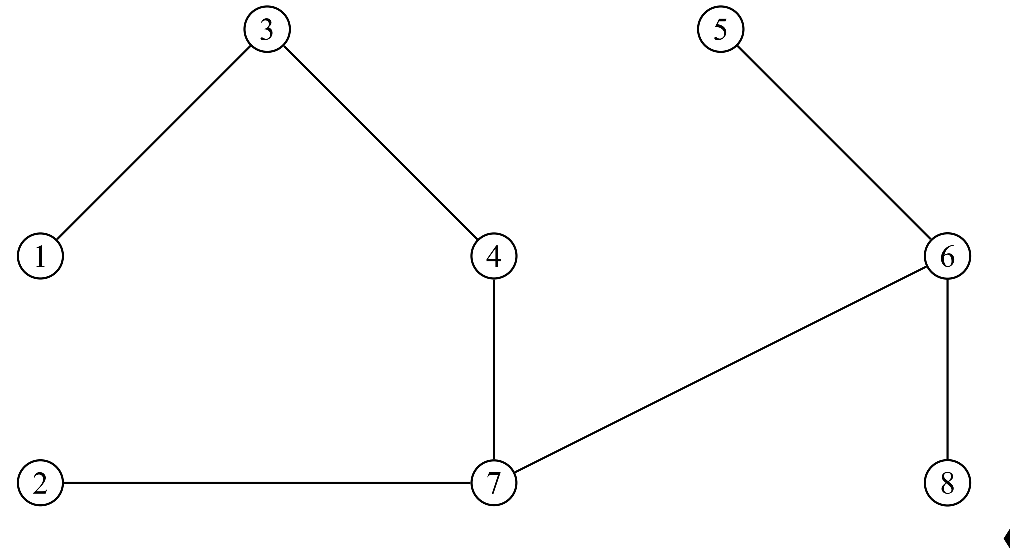
$\{(3,4)_1, (4,7)_1, (5,6)_1, (2,4)_2, (2,7)_2, (1,3)_3, (7,6)_3, (4,5)_4, (4,6)_6, (1,4)_7, (6,8)_7,$
$(1,2)_9, (7,8)_{16}\}$

Teniendo en cuenta que un árbol de n vértices contiene $n-1$ aristas, se tiene que:

- Se toma como C el conjunto $C = \{\{3,4\}\}$
- Se toma la arista $\{4,7\}$, la cual al no cerrar circuito se añade a C, y puesto que $\text{card}\,(C) = 2 \neq \text{card}\,(V(G)) - 1 = 7$.
- Se toma una nueva arista, $\{5,6\}$, la cual tampoco cierra circuito, por lo que se añade a C; dado que $\text{card}\,(C) = 3 \neq 7$.
- Se toma una nueva arista, $\{2,4\}$, la cual tampoco cierra circuito, por lo que se añade a C.
- Se toma una nueva arista, $\{2,7\}$, la cual cierra circuito, por lo que no se incluye en C.
- Se toma una nueva arista, $\{1,3\}$, la cual no cierra circuito, por lo que se añade a C.
- Se toma una nueva arista, $\{6,7\}$, la cual tampoco cierra circuito, por lo que se añade a C.
- Se toma una nueva arista, $\{4,5\}$, la cual cierra circuito, por lo que no se incluye en C.

- Se toma una nueva arista, $\{4, 6\}$, la cual cierra circuito, por lo que no se incluye en C.
- Se toma una nueva arista, $\{1, 7\}$, la cual cierra circuito, por lo que no se incluye en C.
- Se toma una nueva arista, $\{6, 8\}$, la cual tampoco cierra circuito, por lo que se añade a C. El algoritmo finaliza puesto que $\text{card}\,(C) = 7 = \text{card}\,(V(G)) - 1$.

El conector mínimo viene dado por el subgrafo $(V(G), C)$, $C = \{\{3, 4\}, \{4, 7\}, \{5, 6\}, \{2, 4\}, \{1, 3\}, \{6, 7\}, \{6, 8\}\}$ y cuya representación viene dada por:



♦

A continuación se muestra otro ejemplo en el que se utiliza un método matricial más adecuado para ser implementado en el ordenador.

Ejemplo 2.2.3
Sea el grafo G cuya matriz de adyacencia y matriz de pesos vienen dadas por:

$$Ad(G) = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} \qquad \Omega(G) = \begin{bmatrix} 0 & 5 & 1 & \infty & \infty & \infty \\ 5 & 0 & 2 & 1 & 5 & \infty \\ 1 & 2 & 0 & 1 & 2 & \infty \\ \infty & 1 & 1 & 0 & 2 & 5 \\ \infty & 5 & 2 & 2 & 0 & 1 \\ \infty & \infty & \infty & 5 & 1 & 0 \end{bmatrix}$$

Encontrar un conector mínimo usando el método matricial.

En primer lugar, consideramos el conjunto de aristas ordenadas de modo creciente en peso. Así, tenemos el conjunto de aristas ordenado

$$E(G) = \{\{1, 3\}, \{2, 4\}, \{3, 4\}, \{5, 6\}, \{2, 3\}, \{3, 5\}, \{4, 5\}, \{1, 2\}, \{2, 5\}, \{4, 6\}\}$$

Sea G_0 el grafo nulo de seis vértices, se toma la primera arista de $E(G)$ ordenado, y

sea D_1 la matriz de adyacencia del grafo unión de G_0 con el grafo de vértices $\{1, 3\}$ y arista $\{(1, 3)\}$ (grafo al que llamaremos G_1), la cual viene dada por:

$$D_1 = Ad(G_1) = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Las potencias de D_1 resultan:

$$D_1^2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad D_1^3 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$D_1^4 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad D_1^5 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Así, la matriz $\sum_{r=0}^{5} D_1^r$ resulta:

$$\sum_{r=0}^{5} D_1^r = \begin{bmatrix} 3 & 0 & 3 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 3 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

El grafo G_1 es no conexo, pues la matriz $\sum_{r=0}^{5} D_1^r$ contiene ceros. Por tanto, a continuación tomamos la arista $\{2, 4\}$, la cual no cierra circuito al añadirla a G_1, pues entre los vértices 2 y 4 no existe ningún camino (véase la matriz $\sum_{r=0}^{5} D_1^r$ y obsérvese que el elemento (2,4) es cero). Sea G_2 el grafo unión de G_1 con $(\{2, 4\}, \{\{2, 4\}\})$, esto es el grafo que tiene por vértices a los nodos 2 y 4 y por única arista a $\{2, 4\}$.

Sea D_2 la matriz de adyacencia de G_2, la cual viene dada por:

$$D_2 = Ad(G_2) = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Las potencias de D_1 resultan:

$$D_2^2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad D_2^3 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$D_2^4 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad D_2^5 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Así, la matriz $\sum_{r=0}^{5} D_2^r$ resulta:

$$\sum_{r=0}^{5} D_2^r = \begin{bmatrix} 3 & 0 & 3 & 0 & 0 & 0 \\ 0 & 3 & 0 & 3 & 0 & 0 \\ 3 & 0 & 3 & 0 & 0 & 0 \\ 0 & 3 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

El grafo G_2 tampoco resulta conexo, pues la matriz $\sum_{r=0}^{5} D_2^r$ contiene ceros.
A continuación se toma la siguiente arista, la cual corresponde a la $\{3, 4\}$. Sea el grafo $G_3 = G_2 \cup (\{3, 4\}, \{\{3, 4\}\})$, cuya matriz de adyacencia D_3 viene dada por

$$D_3 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

y sus potencias por

$$D_3^2 = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad D_3^3 = \begin{bmatrix} 0 & 1 & 2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 2 & 0 & 0 \\ 2 & 0 & 0 & 3 & 0 & 0 \\ 0 & 2 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$D_3^4 = \begin{bmatrix} 2 & 0 & 0 & 3 & 0 & 0 \\ 0 & 2 & 3 & 0 & 0 & 0 \\ 0 & 3 & 5 & 0 & 0 & 0 \\ 3 & 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad D_3^5 = \begin{bmatrix} 0 & 3 & 5 & 0 & 0 & 0 \\ 3 & 0 & 0 & 5 & 0 & 0 \\ 5 & 0 & 0 & 8 & 0 & 0 \\ 0 & 5 & 8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Así, la matriz $\sum_{r=0}^{5} D_3^r$ resulta:

$$\sum_{r=0}^{5} D_3^r = \begin{bmatrix} 4 & 4 & 8 & 4 & 0 & 0 \\ 4 & 4 & 4 & 8 & 0 & 0 \\ 8 & 4 & 8 & 12 & 0 & 0 \\ 4 & 8 & 12 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

El grafo G_3 tampoco es conexo, pues la matriz $\sum_{r=0}^{5} D_3^r$ contiene ceros.

A continuación se toma la siguiente arista, la cual corresponde a la $\{5, 6\}$. Sea el grafo $G_4 = G_3 \cup (\{5, 6\}, \{\{5, 6\}\})$, cuya matriz de adyacencia D_4 viene dada por

$$D_4 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

y sus potencias por

$$D_4^2 = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \qquad D_4^3 = \begin{bmatrix} 0 & 1 & 2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 2 & 0 & 0 \\ 2 & 0 & 0 & 3 & 0 & 0 \\ 0 & 2 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$D_4^4 = \begin{bmatrix} 2 & 0 & 0 & 3 & 0 & 0 \\ 0 & 2 & 3 & 0 & 0 & 0 \\ 0 & 3 & 5 & 0 & 0 & 0 \\ 3 & 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \qquad D_4^5 = \begin{bmatrix} 0 & 3 & 5 & 0 & 0 & 0 \\ 3 & 0 & 0 & 5 & 0 & 0 \\ 5 & 0 & 0 & 8 & 0 & 0 \\ 0 & 5 & 8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Así, la matriz $\sum_{r=0}^{5} D_4^r$ resulta:

$$\sum_{r=0}^{5} D_4^r = \begin{bmatrix} 4 & 4 & 8 & 4 & 0 & 0 \\ 4 & 4 & 4 & 8 & 0 & 0 \\ 8 & 4 & 8 & 12 & 0 & 0 \\ 4 & 8 & 12 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 0 & 3 & 3 \end{bmatrix}$$

El grafo resultante tampoco es conexo, pues la matriz $\sum_{r=0}^{5} D_4^r$ sigue conteniendo ceros.
La siguiente arista es $\{2, 3\}$. Dicha arista no puede ser incluida por cerrar circuito (el elemento (2,3) de la matriz $\sum_{r=0}^{5} D_4^r$ es 4 y por tanto existen caminos de 2 a 3 que no contienen la arista $\{2, 3\}$, por tanto, su inclusión cierra circuito). Sí es posible incluir $\{3, 5\}$; así resulta $G_5 = G_3 \cup (\{3, 5\}, \{\{3, 5\}\})$, cuya matriz de adyacencia D_5 viene dada por

$$D_5 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

y sus potencias por

$$D_5^2 = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 3 & 0 & 0 & 0 \\ 1 & 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}, \qquad D_5^3 = \begin{bmatrix} 0 & 1 & 3 & 0 & 0 & 1 \\ 1 & 0 & 0 & 2 & 1 & 0 \\ 3 & 0 & 0 & 4 & 4 & 0 \\ 0 & 2 & 4 & 0 & 0 & 1 \\ 0 & 1 & 4 & 0 & 0 & 2 \\ 1 & 0 & 0 & 1 & 2 & 0 \end{bmatrix}$$

$$D_5^4 = \begin{bmatrix} 3 & 0 & 0 & 4 & 4 & 0 \\ 0 & 2 & 4 & 0 & 0 & 1 \\ 0 & 4 & 11 & 0 & 0 & 4 \\ 4 & 0 & 0 & 6 & 5 & 0 \\ 4 & 0 & 0 & 5 & 6 & 0 \\ 0 & 1 & 4 & 0 & 0 & 2 \end{bmatrix}, \qquad D_5^5 = \begin{bmatrix} 0 & 4 & 11 & 0 & 0 & 4 \\ 4 & 0 & 0 & 6 & 5 & 0 \\ 11 & 0 & 0 & 15 & 15 & 0 \\ 0 & 6 & 15 & 0 & 0 & 5 \\ 0 & 5 & 15 & 0 & 0 & 6 \\ 4 & 0 & 0 & 5 & 6 & 0 \end{bmatrix}$$

Así, la matriz $\sum\limits_{r=0}^{5} D_5^r$ resulta:

$$\sum_{r=0}^{5} D_5^r = \begin{bmatrix} 5 & 5 & 15 & 5 & 5 & 5 \\ 5 & 4 & 5 & 9 & 6 & 1 \\ 15 & 5 & 15 & 20 & 20 & 5 \\ 5 & 9 & 20 & 9 & 6 & 6 \\ 5 & 6 & 20 & 6 & 9 & 9 \\ 5 & 1 & 5 & 6 & 9 & 4 \end{bmatrix}$$

El grafo G_5 es conexo puesto que la matriz no contiene ceros, por tanto el conector mínimo es

$$G_5 = (V(G), \{\{1,3\}, \{2,4\}, \{3,4\}, \{3,5\}, \{5,6\}\})$$

♦

2.2.2. El problema del camino más corto

Otro problema clásico en teoría de grafos ponderados es el llamado problema del camino más corto o de peso mínimo, el cual consiste en dados dos vértices de un grafo, encontrar entre las trayectorias que unen dichos vértices una de peso mínimo. Para ello se establecen las siguientes definiciones.

Definición 2.2.3
Sea G un grafo ponderado, y sea una trayectoria $v_1 \to v_2 \to \ldots \to v_k$ entre los vértices v_1 y v_k. Llamamos **peso** (o **costo**) de dicha trayectoria al valor $\omega(v_1, v_2, \ldots, v_k) = \sum\limits_{i=1}^{k-1} \omega(v_i, v_{i+1})$.

Definición 2.2.4
Sea G un grafo conexo y sean $i, j \in V(G)$. Diremos que la trayectoria $i = v_1 \to v_2 \to \ldots \to v_k = j$ es de peso mínimo si para cualquier otra trayectoria $i = w_1 \to w_2 \to \ldots \to w_m = j$ se verifica que $\omega(v_1, v_2, \ldots, v_k) \leq \omega(w_1, w_2, \ldots, w_m)$.

Definición 2.2.5
Llamamos **problema de la trayectoria más corta** al problema consistente en: dado un grafo conexo G y dados dos vértices $i, j \in V(G)$, encontrar una trayectoria que conecte i con j de peso mínimo.

Para resolver este problema se utiliza el llamado principio de minimalidad de Bellman, el cual podemos enunciar como sigue:

Principio de minimalidad de Bellman

Sea G un grafo conexo y sea un vértice $i \in V(G)$. Sea un vértice $k \in V(G)$ alcanzable desde i mediante una trayectoria de costo $l(k)$. Si j es un vértice adyacente a k de modo que existe una trayectoria de costo $l(j)$, que conecta i con j y no contiene a k, entonces, existe una trayectoria que conecta i con k cuyo costo es menor o igual que $\min\{l(k), l(j) + \omega(\{j,k\})\}$.

En base a este principio, Dijkstra propuso un algoritmo capaz de proporcionar una trayectoria de costo mínimo entre dos vértices L y S de un grafo conexo, conocido como algoritmo de Dijkstra, el cual podemos enunciar como:

Algoritmo de Dikjstra

1. A cada vértice del grafo se le asigna una etiqueta $l(i)$, definida como $l(L) = 0$ y $l(i) = \infty$ si $i \neq L$.

2. Se localiza un vértice i con etiqueta $l(i)$ mínima (si hay varios de tales vértices se toma uno cualquiera de ellos).

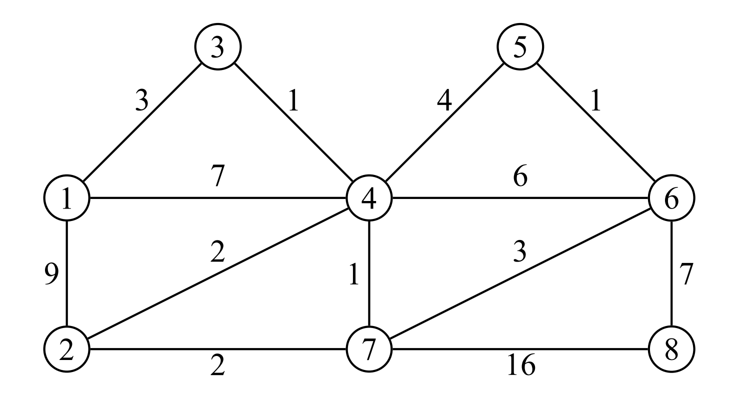
   - Si $i = S$, entonces la longitud del camino mas corto entre los vértices L y S es $l(S)$. En esta situación finaliza el algoritmo.
   - Si $i \neq S$, entonces se pasa a la etapa siguiente.

3. Para cada vértice j adyacente a i se determina una nueva etiqueta como $l(j) = \min\{l(j), l(i) + \omega(\{i,j\})\}$.

4. Se elimina del grafo el vértice i y las aristas que lo contienen.

5. Se vuelve al paso dos.

Ejemplo 2.2.4
Dado el grafo representado en la figura:



Encontrar el camino más corto entre uno y seis.

La solución se obtiene mediante el algoritmo de Dijkstra el cual se puede expresar en forma tabular como

	1	2	3	4	5	6	7	8
	0	∞	∞	∞	∞	∞	∞	∞
1	-	9	3	7	∞	∞	∞	∞
3	-	9	-	4	∞	∞	∞	∞
4	-	6	-	-	8	10	5	∞
7	-	6	-	-	8	8	-	21
2	-	-	-	-	8	8	-	21

El mínimo costo necesario para ir de 1 a 6 resulta ser de 8 unidades, lo cual se logra a través de la trayectoria

$$1 \to 3 \to 4 \to 7 \to 6$$

Nota: La tabla anterior se ha construido del modo siguiente. En cada momento las columnas representan las etiquetas de los vértices que quedan, indicándose mediante el signos $-$ cuando el vértice ha sido eliminado.

Puesto que en nuestro caso $L = 1$ y $S = 6$, inicialmente se asigna $l(L) = 0$ y $L(i) = \infty$ si $i \neq L$ resultando

1	2	3	4	5	6	7	8
0	∞	∞	∞	∞	∞	∞	∞

El vértice de etiqueta mínima resulta ser el 1. Por tanto, para los vértices i adyacentes a 1 se cambia la etiqueta $l(i)$ por $\text{mín}\{l(i), l(1) + \omega(\{1, i\})\}$ permaneciendo el resto inalterables, eliminado finalmente el vértice 1. Así se asigna para $l(2)$ el valor $\text{mín}\{l(2), l(1)+\omega(\{1, 2\})\}$, de donde $l(2) = \text{mín}\{+\infty, 0+9\} = 9$. Análogamente para los vértices 3 y 4, se tiene $l(3) = 3$, $l(4) = 7$.

Si a la izquierda de la tabla se ha escrito el vértice que tenía etiqueta mínima en la etapa anterior, resulta

	1	2	3	4	5	6	7	8
	0	∞	∞	∞	∞	∞	∞	∞
1	-	9	3	7	∞	∞	∞	∞

En este momento el vértice de menor etiqueta es 3, que no coincide con S; por tanto repitiendo el proceso se tiene

	1	2	3	4	5	6	7	8
	0	∞	∞	∞	∞	∞	∞	∞
1	-	9	3	7	∞	∞	∞	∞
3	-	9	-	4	∞	∞	∞	∞

y así sucesivamente hasta que la etiqueta mínima corresponde al vértice de destino (pudiendo haber o no otros con el mismo costo). En este momento el costo mínimo es el marcado por la etiqueta $l(6) = 8$. La trayectoria de costo mínimo se obtiene observando desde qué vértice k se ha conseguido el cambio de la etiqueta $l(6)$, desde qué vértice j se ha producido el cambio de la etiqueta $l(k)$, y así sucesivamente hasta llegar al vértice 1:

	1	2	3	4	5	6	7	8
	0	∞	∞	∞	∞	∞	∞	∞
1	-	9	3	7	∞	∞	∞	∞
3	-	9	-	4	∞	∞	∞	∞
4	-	6	-	-	8	10	5	∞
7	-	6	-	-	8	8	-	21
2	-	-	-	-	8	8	-	21

así 6 viene de 7, 7 viene de 4, 4 viene de 3 y 3 viene de 1. Por tanto, la trayectoria de mínimo costo encontrada es

$$1 \to 3 \to 4 \to 7 \to 6$$

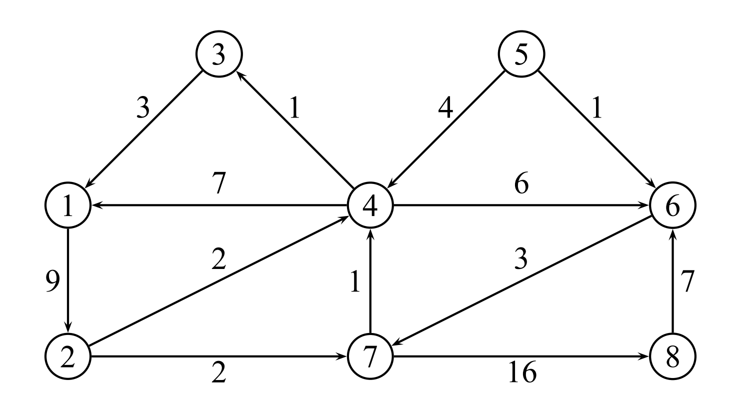
. ♦

2.3. Optimización en dígrafos

A continuación se aborda el estudio de algunos problemas de optimización, los cuales pueden ser modelados a partir de la teoría de dígrafos ponderados.

Definición 2.3.1
Sea D un dígrafo. Una **ponderación** en D es una aplicación $\omega : A(G) \to \mathbb{R}^+ \cup \{0\}$. En ese caso, el par (D, ω) recibe el nombre de **dígrafo ponderado** y el valor $\omega(e)$, donde $e \in A(G)$, el de **peso** (o **costo**) del arco e.

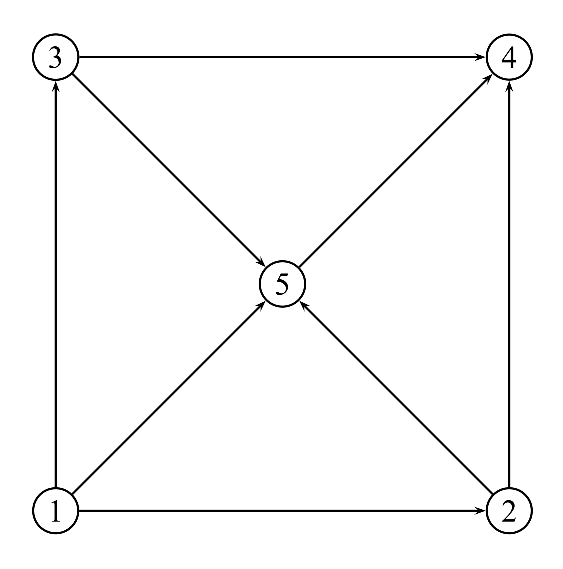
Ejemplo 2.3.1
A continuación se da la representación de un dígrafo ponderado.



♦

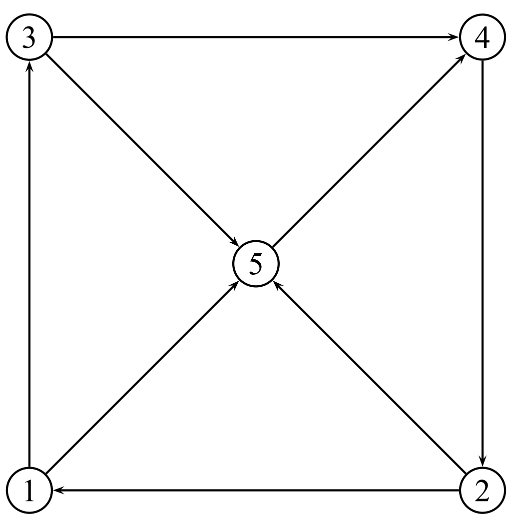
Definición 2.3.2
Sea D un dígrafo débilmente conexo. Diremos que D es **acíclico** si no contiene ningún circuito orientado.

Ejemplo 2.3.2
El dígrafo cuya representación aparece a continuación es un grafo acíclico, pues no contiene circuitos orientados.



♦

Ejemplo 2.3.3
El dígrafo cuya representación aparece a continuación no es un grafo acíclico pues contiene, entre otros, los circuitos orientados $1 \to 3 \to 4 \to 2 \to 1$ y $5 \to 4 \to 2 \to 5$.



♦

Definición 2.3.3
Sea D un dígrafo débilmente conexo. Diremos que un vértice k es **vértice inicial** o **vértice fuente** si no existen arcos de la forma (i, k) donde $i \in A(D) - \{k\}$.

Definición 2.3.4
Sea D un dígrafo débilmente conexo. Diremos que un vértice k es **vértice final** o **vértice sumidero** si no existen arcos de la forma (k, i) donde $i \in A(D) - \{k\}$.

Ejemplo 2.3.4
En el dígrafo correspondiente al ejemplo de grafo acíclico, el vértice 1 es un vértice inicial y el vértice 4 un vértice final. ♦

Definición 2.3.5
Llamamos **red** a un dígrafo D con dos ponderaciones llamadas **capacidades** y **flujos**, que representaremos como c y f, las cuales satisfacen $\forall(i, j) \in A(D)$ $c_{i,j} = c(i, j) \geq 0$ y $0 \leq f_{i,j} = f(i, j) \leq c_{i,j}$.

En este libro manejaremos únicamente redes acíclicas. En este caso es posible definir para cada par de vértices las capacidades $(i, j) \in V(D) \times V(D)$;

$$c_{i,j} = \begin{cases} c(i,j) & (i,j) \in A(D) \\ -c(i,j) & (j,i) \in A(D) \\ 0 & \{i,j\} \notin E(D) \end{cases}$$

y los flujos

$$f_{i,j} = \begin{cases} f(i,j) & (i,j) \in A(D) \\ -f(i,j) & (j,i) \in A(D) \\ 0 & \{i,j\} \notin E(D) \end{cases}$$

siendo en ambos casos $E(D)$ el conjunto de aristas del grafo subyacente a D.

Finalmente indicar que en un dígrafo acíclico D, es posible definir una relación de orden dada por $\forall i, j \in V(D)$, $i \leq j$ si, y sólo si, existe en caso de ser $i \neq j$ una trayectoria dirigida que une i con j (compruébese que esta relación es de orden).

2.3.1. El problema de la trayectoria crítica

Muchos proyectos pueden ser divididos en una serie de tareas de modo que para poder realizar unas de ellas deben estar finalizadas otras (una o varias), pudiéndose dar por finalizado el trabajo únicamente cuando están acabadas todas las tareas. Las tareas pueden modelarse como arcos en un dígrafo, asignándose como peso el tiempo que cuesta realizar dicha tarea. Como nodos representaremos los requisitos necesarios (tareas que deben estar finalizadas) para emprender una nueva. Las tareas que no necesitan ningún prerrequisito partirán del vértice inicial, y aquellas tareas no necesarias como punto de partida para ninguna otra, finalizarán en un vértice final.

Un problema de gran interés es la evaluación del tiempo mínimo en que puede ser ejecutado un proyecto. Este problema puede ser modelado mediante el llamado problema de la trayectoria crítica.

Definición 2.3.6
Sea D un dígrafo acíclico ponderado. Diremos que una **trayectoria** es **crítica** si es una trayectoria que conecta un vértice inicial con un vértice final de modo que su costo es máximo.

El problema de la trayectoria crítica puede resolverse mediante el algoritmo PERT (técnica de revisión y evaluación de programas). Dicho algoritmo puede esquematizarse como

Algoritmo PERT

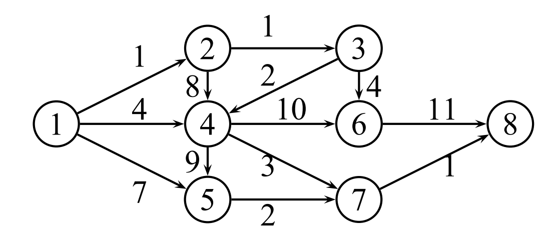
1. Etiquetar todos los vértices como cero: $l(i) = 0$, $\forall i \in V(D)$.

2. Tomar un vértice inicial i. Para todos los vértices adyacentes j asignar a $l(j)$ el valor $\max\{l(j), l(i) + l(i,j)\}$. Eliminar el vértice i y los arcos que parten de él.

3. Si todos los vértices restantes son vértices finales, el algoritmo finaliza siendo el costo de la trayectoria crítica la mayor de las etiquetas de los vértices finales. Si existen vértices no finales ir al paso 2.

El algoritmo PERT puede aplicarse en forma tabular tal como se muestra en el ejemplo.

Ejemplo 2.3.5
Obtener una trayectoria crítica en el dígrafo cuya representación viene dada en la siguiente figura.



El problema se resuelve mediante el algoritmo del PERT el cual en forma tabular resulta:

	1	2	3	4	5	6	7	8
	0	0	0	0	0	0	0	0
1	-	1	0	4	7	0	0	0
2	-	-	2	9	7	0	0	0
3	-	-	-	9	7	6	0	0
4	-	-	-	-	18	19	12	0
5	-	-	-	-	-	19	20	0
6	-	-	-	-	-	-	20	30
7	-	-	-	-	-	-	-	30

Por tanto, la trayectoria crítica tiene un coste de 30. Una trayectoria crítica es la siguiente:

$$1 \to 2 \to 4 \to 6 \to 8$$

♦

2.3.2. El problema del flujo máximo en una red

En este epígrafe se aborda el estudio del flujo máximo que puede pasar a través de una red, o el equivalente consistente en investigar en cuánto se puede incrementar el flujo que atraviesa una red. Para ello, sea D un dígrafo acíclico sobre el cual se han

definido unas capacidades dadas por una matriz C, cuyos elementos denotaremos por $c_{i,j}$ sobre la cual circulan unos flujos dados por una matriz F, cuyos elementos denotaremos por $f_{i,j}$. Las matrices C y F son matrices antisimétricas.

En este estudio consideraremos que la red contiene una única fuente y un único sumidero (en otro caso, se puede reducir fácilmente el problema a éste). Sea L el vértice fuente y S el vértice sumidero. Entonces el flujo f que atraviesa la red viene dado por $f = \sum_{i=1}^{n} f_{L,i} = \sum_{i=1}^{n} f_{i,S}$. En el resto de nodos que denominaremos intermedios debe cumplirse la ley de Kirchoff (flujo entrante igual a flujo saliente) lo que implica la condición $\sum_{j=0}^{n} f_{i,j} = 0, \forall i \in \{1, \ldots, n\} - \{L, S\}$.

El método que seguiremos para buscar un incremento de flujo en la red consistirá en investigar trayectorias entre L y S, a través de las cuales es posible hacer llegar una cantidad adicional de flujo al nodo n. Para poder enviar desde un nodo una cantidad adicional de flujo, siempre sin exceder la capacidad, hay dos procedimientos: uno, hacer llegar más flujo al nodo, y otro, enviar menos cantidad de flujo a otro nodo. Basándose en este principio, Ford y Fulkerson construyeron un algoritmo capaz de encontrar una trayectoria de L a S, caso de ser posible, a través de la cual incrementar el flujo que circula por la red.

Algoritmo de Ford-Fulkerson

1. Etiquetar L como $\triangle_1 = +\infty$, dejando el resto de vértices como no etiquetados.

2. Buscar un vértice i etiquetado no analizado. Procédase del siguiente modo: para cada vértice j no etiquetado adyacente a i, si $0 \leq f_{i,j} < c_{i,j}$, calcular:

   $$\triangle_{i,j} = c_{i,j} - f_{i,j} \quad y \quad \triangle_j = \text{mín}\{\triangle_i, \triangle_{i,j}\}$$

   y etiquétese el vértice j como $(\triangle_j, i^+)$. En caso de ser $f_{i,j} < 0$, calcular

   $$\triangle_{i,j} = -f_{i,j} \quad y \quad \triangle_j = \text{mín}\{\triangle_i, \triangle_{i,j}\}$$

   y etiquétese j como $(\triangle_j, i^-)$. Marcar el vértice i como analizado.

   Si no existe tal vértice i, el flujo actual es el flujo máximo, con lo que finaliza el algoritmo.

3. Repetir el paso 2 hasta que se alcanza el punto S.

4. Trazar el camino de L a S utilizando las etiquetas, e incrementar el flujo a través de dicho camino en $\triangle_S$ unidades.

5. A partir del camino construido anteriormente, calcular el flujo incrementado como $f = f + \triangle_S$.

6. Eliminar las marcas de etiquetado y analizado de los vértices y volver al paso 1 para proceder a una nueva iteración.

En este algoritmo debe entenderse:

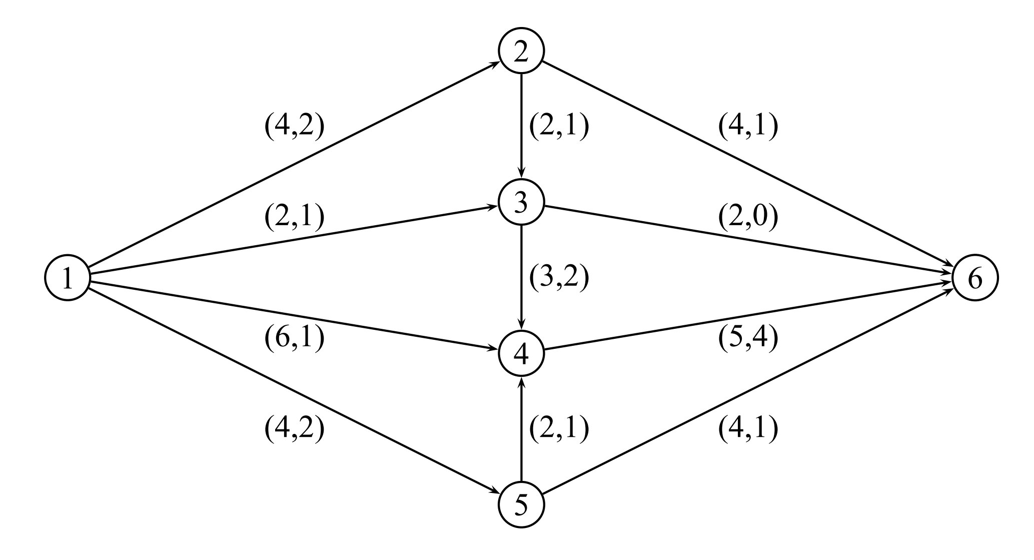
- Por vértice analizado en cada iteración se entenderá aquél para el cual se ha estudiado en cuanto se puede incrementar el flujo en los vértices adyacentes según el paso 2.
- Dos vértices son adyacentes si lo son en el grafo subyacente a D.

Ejemplo 2.3.6
Dado la red cuyas matrices de capacidades y de flujos vienen dadas respectivamente por (se da únicamente la parte superior de ambas matrices):

$$C = \begin{bmatrix} 0 & 4 & 2 & 6 & 4 & 0 \\ . & 0 & 2 & 0 & 0 & 4 \\ . & . & 0 & 3 & 0 & 2 \\ . & . & . & 0 & -2 & 5 \\ . & . & . & . & 0 & 4 \\ . & . & . & . & . & 0 \end{bmatrix} \qquad F = \begin{bmatrix} 0 & 2 & 1 & 1 & 2 & 0 \\ . & 0 & 1 & 0 & 0 & 1 \\ . & . & 0 & 2 & 0 & 0 \\ . & . & . & 0 & -1 & 4 \\ . & . & . & . & 0 & 1 \\ . & . & . & . & . & 0 \end{bmatrix}$$

determinar el flujo máximo que puede atravesar dicha red (utilizar el método de Ford-Fulkerson en su forma tabular).

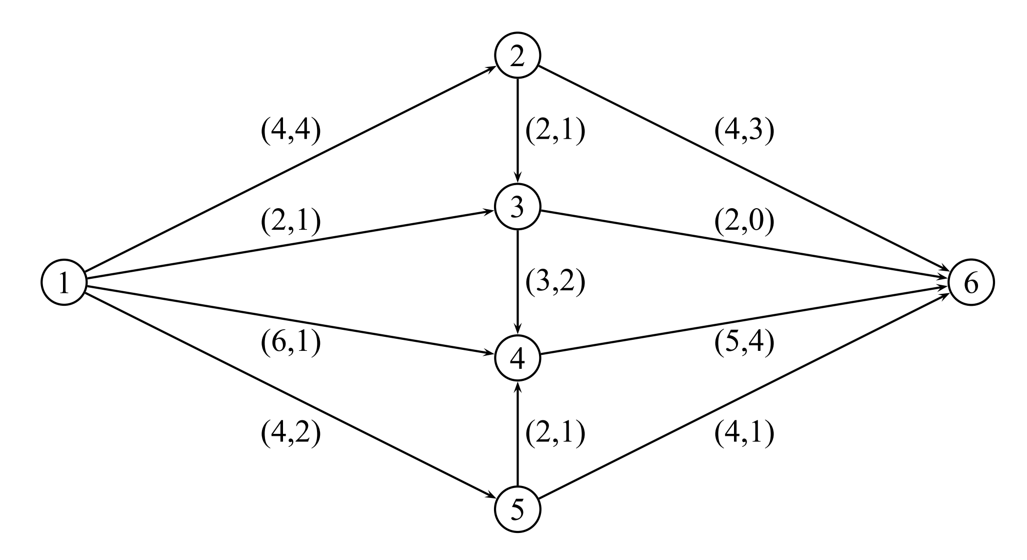
En primer lugar, se tiene que el grafo puede ser representado como



El flujo inicial es $f = \sum_{j=2}^{6} f_{1,j} = 6$. Aplicando el algoritmo de Ford-Fulkerson se tiene:

	1	2	3	4	5	6
	$+\infty$	-	-	-	-	-
1	$+\infty$	$(2, 1^+)$	$(1, 1^+)$	$(5, 1^+)$	$(2, 1^+)$	-
2	$+\infty$	$(2, 1^+)$	$(1, 1^+)$	$(5, 1^+)$	$(2, 1^+)$	$\triangle_{2,6} = 3$ $\triangle_6 = (2, 2^+)$

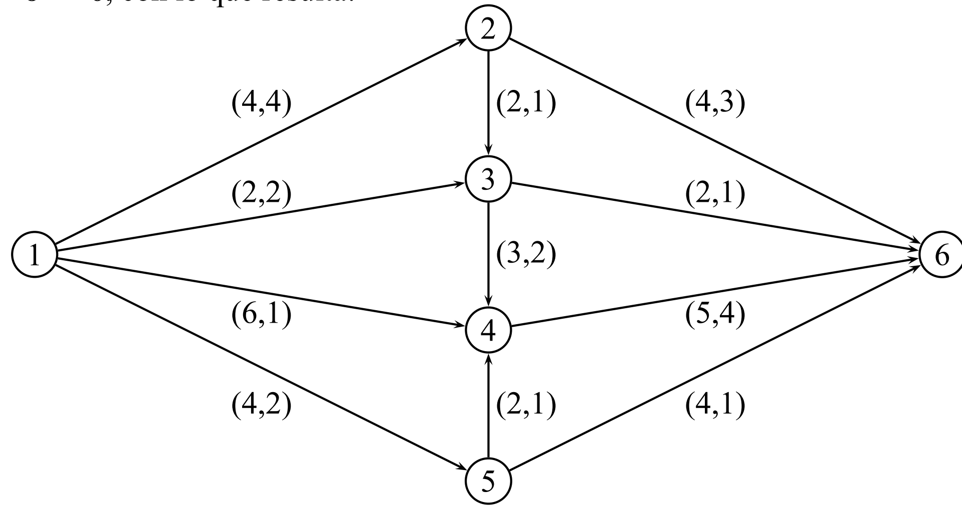
Por tanto ,la red admite un incremento de flujo de dos unidades a través del camino $1 \to 2 \to 6$. Incrementando el flujo actual en dos unidades a través del camino resulta:



resultando el flujo actual $f = 8$ unidades.

	1	2	3	4	5	6
	$+\infty$	-	-	-	-	-
1	$+\infty$	-	$(1, 1^+)$	$(5, 1^+)$	$(2, 1^+)$	-
3	$+\infty$	$\triangle_{3,2} = -1$ $(1, 3^-)$	$(1, 1^+)$	$(5, 1^+)$	$(2, 1^+)$	$\triangle_{3,6} = 2$ $(1, 3^+)$

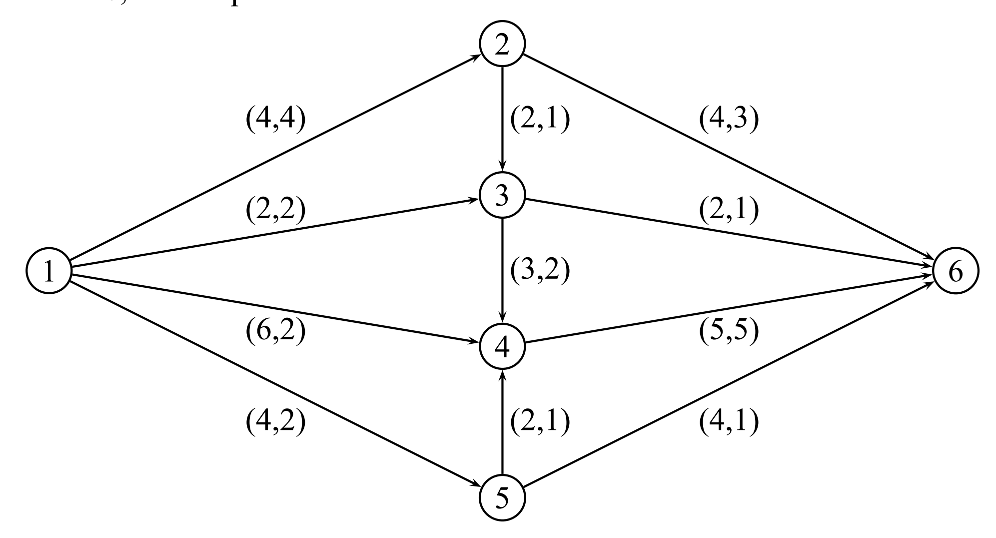
Por tanto, la red admite un incremento de flujo de una unidad a través del camino $1 \to 3 \to 6$, con lo que resulta:



de donde el flujo actual es $f = 9$ unidades.

	1	2	3	4	5	6
	$+\infty$	-	-	-	-	-
1	$+\infty$	-	-	$(5, 1^+)$	$(2, 1^+)$	-
4	$+\infty$	-	$\triangle_{4,3} = -2$ $(2, 4^-)$	$(5, 1^+)$	$(2, 1^+)$	$\triangle_{4,6} = 1$ $(1, 4^+)$

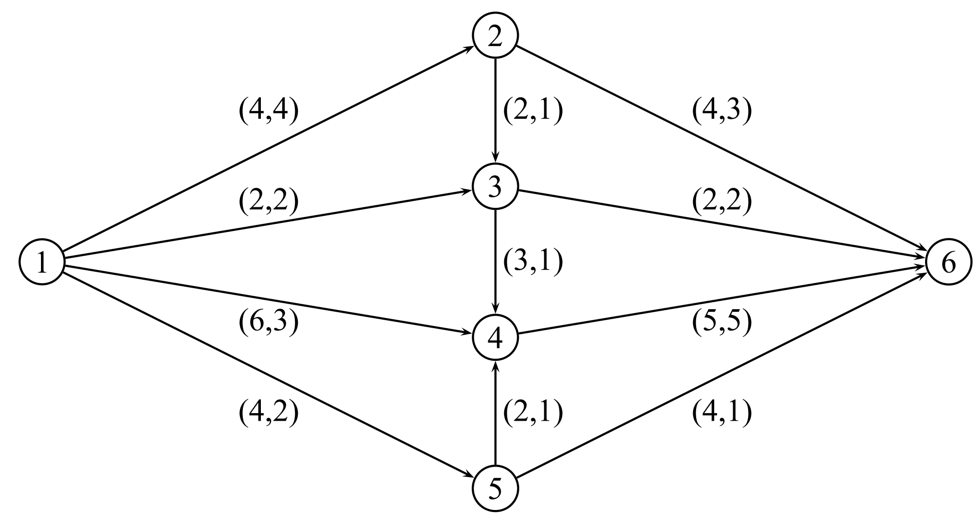
Por tanto, la red admite un incremento de flujo de una unidad a través del camino $1 \to 4 \to 6$, con lo que resulta:



El flujo actual es de diez unidades. Aplicando nuevamente el algoritmo se tiene:

	1	2	3	4	5	6
	$+\infty$	-	-	-	-	-
1	$+\infty$	-	-	$(4, 1^+)$	$(2, 1^+)$	-
4	$+\infty$	-	$\triangle_{4,3} = -2$ $(2, 4^-)$	$(4, 1^+)$	$(2, 1^+)$	-
3	$+\infty$	$\triangle_{3,2} = -1$ $(1, 3^-)$	$(2, 4^-)$	$(4, 1^+)$	$(2, 1^+)$	$\triangle_{3,6} = 1$ $(1, 3^+)$

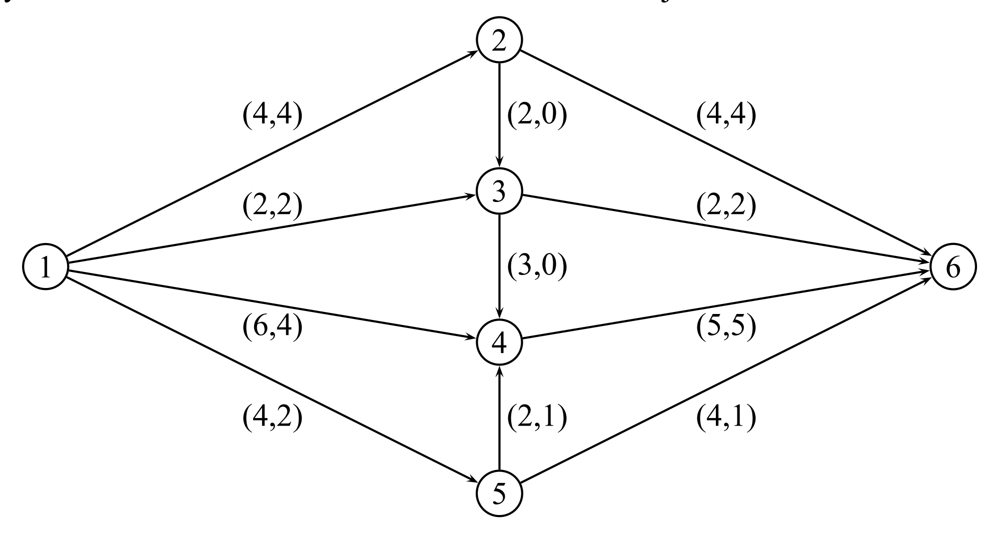
Por tanto, es posible aumentar el flujo de la red en una unidad a través del camino $1 \to 4 \leftarrow 3 \to 6$. Así, resulta:



El flujo actual resulta de once unidades. Aplicando nuevamente el algoritmo se obtiene la tabla de incrementos de flujo:

	1	2	3	4	5	6
	$+\infty$	-	-	-	-	-
1	$+\infty$	-	-	$(3,1^+)$	$(2,1^+)$	-
4	$+\infty$	-	$\triangle_{4,3}=-1$ $(1,4^-)$	$(3,1^+)$	$(2,1^+)$	-
3	$+\infty$	$\triangle_{3,2}=-1$ $(1,3^-)$	$(1,4^-)$	$(3,1^+)$	$(2,1^+)$	- -
2	$+\infty$	$(1,3^-)$	$(1,4^-)$	$(3,1^+)$	$(2,1^+)$	$\triangle_{2,6}=1$ $(1,2^+)$

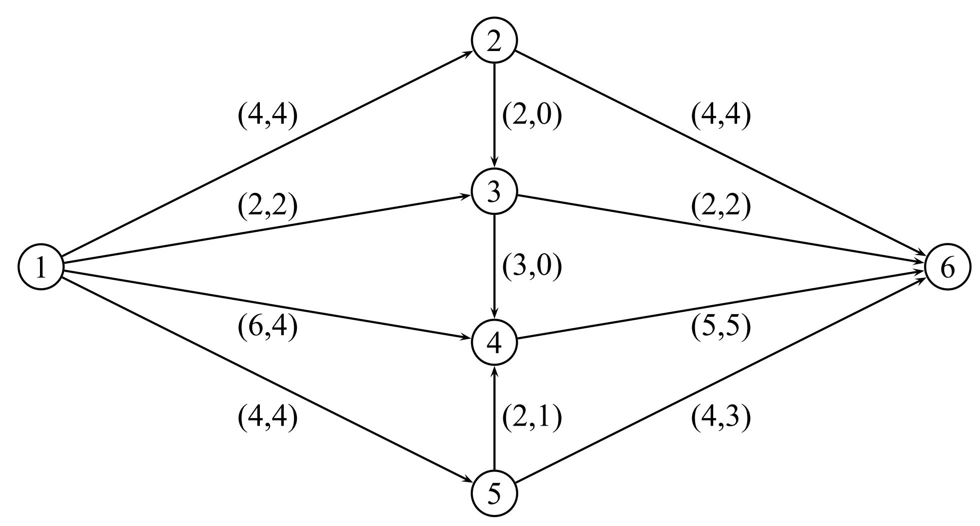
Por tanto, es posible incrementar el flujo en una nueva unidad a través de la trayectoria $1 \to 4 \leftarrow 3 \leftarrow 2 \to 6$. Actualizando los flujos de la red se tiene:



El flujo actual resulta de doce unidades. Aplicando nuevamente el algoritmo se obtiene la tabla de incrementos de flujo:

	1	2	3	4	5	6
	$+\infty$	-	-	-	-	-
1	$+\infty$	-	-	$(2,1^+)$	$(2,1^+)$	-
4	$+\infty$	-	-	$(2,1^+)$	$(2,1^+)$	-
5	$+\infty$	-	-	$(2,1^+)$	$(2,1^+)$	$\triangle_{5,6}=3$ $(2,5^+)$

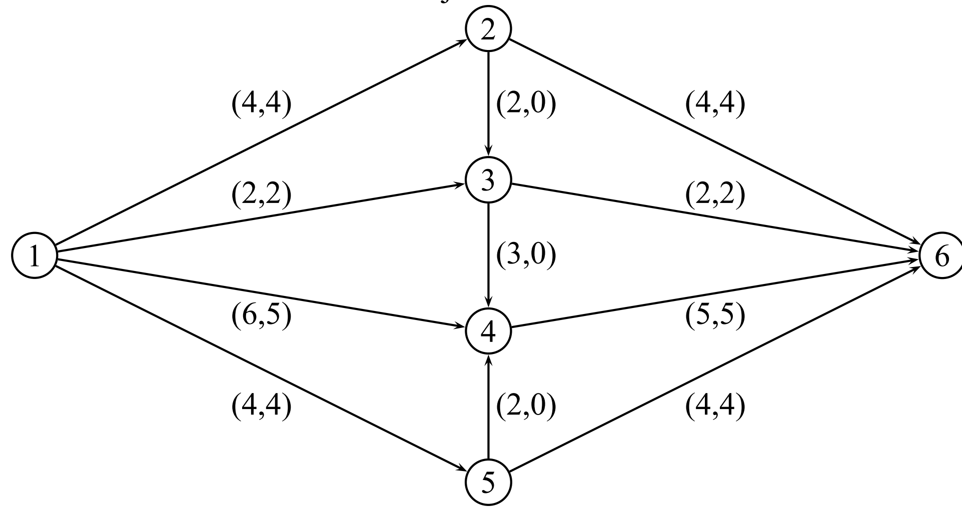
Por tanto, es posible incrementar el flujo dos unidades a través de la trayectoria $1 \to 5 \to 6$. Aplicando de nuevo el algoritmo se tiene:



El flujo actual resulta de catorce unidades. Aplicando nuevamente el algoritmo se obtiene la tabla de incrementos de flujo:

	1	2	3	4	5	6
	$+\infty$	-	-	-	-	-
1	$+\infty$	-	-	$\triangle_{1,4} = 2$ $(2, 1^+)$	-	-
4	$+\infty$	-	-	$(2, 1^+)$	$\triangle_{4,5} = -1$ $(1, 4^-)$	-
5	$+\infty$	-	-	$(2, 1^+)$	$(1, 4^-)$	$\triangle_{5,6} = 1$ $(1, 5^+)$

Por tanto es posible incrementar el flujo en una unidad a través del camino $1 \to 4 \leftarrow 5 \to 6$. Actualizando los flujos de la red se obtiene:



El flujo actual resulta de quince unidades. Aplicando nuevamente el algoritmo de Ford-Fulkerson se tiene la tabla:

	1	2	3	4	5	6
	$+\infty$	-	-	-	-	-
1	$+\infty$	-	-	$\triangle_{1,4} = 1$ $(1, 1^+)$	-	-
4	$+\infty$	-	-	$(1, 1^+)$	-	-

Llegados a este punto no existen nodos etiquetados sin analizar, no habiéndose alcanzado el nodo final, por tanto, según el algoritmo de Ford-Fulkerson ya no son posibles nuevos incrementos de flujo, por lo que el flujo máximo es el flujo actual. ♦

Tema 3

Programación lineal

3.1. Introducción

La programación lineal es un caso particular de un problema más general; la programación matemática que aborda el problema de optimizar una función o funcional $\varphi : \mathcal{X} \rightarrow \mathbb{R}$ definida sobre un cierto subconjunto E de un espacio vectorial real (de dimensión no necesariamente finita) sometida a un conjunto de restricciones $\mathcal{L}$. En el problema expuesto anteriormente, $\mathcal{L}$ puede ser, bien un determinado subconjunto de $\mathbb{R}^k$, bien un subconjunto de un determinado espacio funcional. Atendiendo al tipo de función o funcional a optimizar y al tipo de restricciones, podemos clasificar los problemas de programación matemática como:

- Programación no lineal: optimizar $\varphi(X)$ sometido a $X \in \mathcal{L}$.

- Teoría de control y programación dinámica: optimizar $\varphi(X)$ sometido a $X \in \mathcal{L}$ en caso de ser X un vector de un espacio de dimensión infinita.

A continuación, por su importancia en este curso, se muestra más detalladamente el problema de la programación no lineal como aquél que satisface:

- $\mathcal{L} \subset \mathcal{X} \subset \mathbb{R}^n$.
- $\mathcal{L} = \left\{ X \in \mathcal{X} | g(X) \leq \overrightarrow{b} \right\}$ donde $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$, $\overrightarrow{b} \in \mathbb{R}^m$.
- $\mathcal{L}$ convexo.

Con estas condiciones se excluyen los problemas de infinitas dimensiones, los problemas con un número infinito de restricciones y los problemas discretos (programación entera). Estos últimos serán estudiados posteriormente mediante algoritmos especiales.
Existen tipos especiales en el problema de programación no lineal según sea la función a optimizar o las restricciones; en particular destacamos:

Si $\mathcal{L} = \mathbb{R}^n$ diremos que estamos ante un problema sin restricciones.

Si $\mathcal{X} = \mathbb{R}^n$ y sólo existen restricciones de igualdad diremos que estamos ante un problema clásico de extremos.

Si la función g es un homomorfismo de espacios vectoriales, diremos que estamos ante un problema con restricciones lineales. Este problema tiene dos subclases muy importantes:

Si $\varphi(X) = X^t QX + p^t X$ donde C es una matriz simétrica $n \times n$ y $p \in \mathbb{R}^n$, diremos que estamos ante un problema de programación cuadrática.

Si $\varphi(X) = p^t X$ donde $p \in \mathbb{R}^n$, diremos que estamos ante un problema de programación lineal.

Los problemas de programación dinámica y de control no serán abordados en este curso.

Los métodos de programación requieren algún conocimiento acerca del análisis convexo. Para ello, definimos en primer lugar conjunto convexo.

3.2. Conjuntos convexos

Definición 3.2.1
Sea $K \subseteq \mathbb{R}^n$. Diremos que K es un **conjunto convexo** si $\forall \overrightarrow{x}, \overrightarrow{y} \in K$ y $\forall \lambda \in [0,1]$ se verifica que $\lambda \overrightarrow{x} + (1-\lambda)\overrightarrow{y} \in K$.

Ejemplo 3.2.1
Sea $K = \{(x,y,z) \in \mathbb{R}^3 | x^2 + y^2 + z^2 \leq 1\}$.

Para estudiar si K es convexo, sean dos puntos cualesquiera $(x_1, y_1, z_1), (x_2, y_2, z_2) \in K$ y sea $\lambda \in [0,1]$.

El punto $\lambda(x_1, y_1, z_1) + (1-\lambda)(x_2, y_2, z_2) = (\lambda x_1 + (1-\lambda)x_2, \lambda y_1 + (1-\lambda)y_2,$ $, \lambda z_1 + (1-\lambda)z_2)$ pertenecerá a K si, y sólo si, se verifica que

$$[\lambda x_1 + (1-\lambda)x_2]^2 + [\lambda y_1 + (1-\lambda)y_2]^2 + [\lambda z_1 + (1-\lambda)z_2]^2 \leq 1.$$

Desarrollando el primer miembro se tiene que

$$\lambda^2 x_1^2 + (1-\lambda)^2 x_2^2 + 2\lambda(1-\lambda)x_1x_2 + \lambda^2 y_1^2 + (1-\lambda)^2 y_2^2 + 2\lambda(1-\lambda)y_1y_2 + \\ + \lambda^2 z_1^2 + (1-\lambda)^2 z_2^2 + 2\lambda(1-\lambda)z_1z_2 = \lambda^2(x_1^2 + y_1^2 + z_1^2) + (1-\lambda)^2(x_2^2 + y_2^2 + z_2^2) + \\ 2\lambda(1-\lambda)(x_1x_2 + y_1y_2 + z_1z_2) = (*)$$

Teniendo en cuenta que $x_1^2 + y_1^2 + z_1^2 \leq 1$, $x_2^2 + y_2^2 + z_2^2 \leq 1$ y que $x_1x_2 + y_1y_2 + z_1z_2$ es el producto escalar de los vectores $\overrightarrow{r}_1 = (x_1, y_1, z_1)$,$\overrightarrow{r}_2 = (x_2, y_2, z_2)$ y que $\overrightarrow{r}_1 \cdot \overrightarrow{r}_2 = ||\overrightarrow{r}_1|| \, ||\overrightarrow{r}_2|| \cos\alpha \leq 1$ puesto que $||\overrightarrow{r}_1|| = x_1^2 + y_1^2 + z_1^2 \leq 1$, $||\overrightarrow{r}_2|| = x_2^2 + y_2^2 + z_2^2 \leq 1$ y $\cos\alpha \leq 1$ resulta

$$(*) \leq \lambda^2 + (1-\lambda)^2 + 2\lambda(1-\lambda) = [\lambda + (1-\lambda)]^2 = 1^2 = 1$$

Por tanto, se verifica la condición requerida, lo que implica que el conjunto K es convexo. ♦

Ejemplo 3.2.2
Sea el conjunto $K = \{(x, y) \in \mathbb{R}^2 | x^2 + y^2 \in [1, 2], x \leq 0, y \leq 0\}$. Este conjunto no es convexo, pues los puntos $(1, 0)$,$(0, 1) \in K$; el punto $(\frac{1}{2}, \frac{1}{2})$ pertenece al segmento que une $(1, 0)$ con $(0, 1)$, pues $(\frac{1}{2}, \frac{1}{2}) = \lambda(1, 0) + (1 - \lambda)(0, 1)$ con $\lambda = \frac{1}{2}$, pero dicho punto no pertenece a K, pues $\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 = \frac{1}{2} \notin [1, 2]$. Por tanto, K no es convexo. ♦

Teorema 3.2.1
Sea $K \in \mathbb{R}^N$ un conjunto convexo y sea $A : \mathbb{R}^n \to \mathbb{R}^m$ una transformación afín. Entonces $A(K)$ es convexo en $\mathbb{R}^m$.

Dem:
A transformación afín por tanto, $\forall \overrightarrow{x} \in \mathbb{R}^n$ se tiene que $\exists \overrightarrow{b} \in \mathbb{R}^m$ y $\exists f : \mathbb{R}^n \to \mathbb{R}^m$ lineal de modo que $A(\overrightarrow{x}) = f(\overrightarrow{x}) + \overrightarrow{b}$.

Sean $\overrightarrow{y}_1, \overrightarrow{y}_2 \in f(K)$, y sea $\lambda \in [0, 1]$, entonces $\exists \overrightarrow{x}_1, \overrightarrow{x}_2 \in K$ de modo que $\overrightarrow{y}_1 = A(\overrightarrow{x}_1)$, $\overrightarrow{y}_2 = A(\overrightarrow{x}_2)$, por tanto

$$\begin{aligned} \lambda \overrightarrow{y}_1 + (1 - \lambda)\overrightarrow{y}_2 &= \lambda A(\overrightarrow{x}_1) + (1 - \lambda)A(\overrightarrow{x}_2) = \\ &= \lambda \left(f(\overrightarrow{x}_1) + \overrightarrow{b})+ \right) + (1 - \lambda) \left(f(\overrightarrow{x}_2) + \overrightarrow{b})) + \overrightarrow{b} \right) = \\ &= f(\lambda \overrightarrow{x}_1 + (1 - \lambda)\overrightarrow{x}_2) + \lambda \overrightarrow{b} + (1 - \lambda)\overrightarrow{b} = f(\lambda \overrightarrow{x}_1 + (1 - \lambda)\overrightarrow{b} \end{aligned}$$

y puesto que K es convexo se tiene que $\overrightarrow{x} = \lambda \overrightarrow{x}_1 + (1 - \lambda)\overrightarrow{x}_2 \in K$, así

$$\lambda \overrightarrow{y}_1 + (1 - \lambda)\overrightarrow{y}_2 = f(\overrightarrow{x}) + \overrightarrow{b} \in f(K)$$

y, por tanto, $f(K)$ es convexo. ▼

Teorema 3.2.2
Sea $\{K_i\}_{i \in I}$ una familia de conjuntos convexos de $\mathbb{R}^n$. Entonces $\bigcap_{i \in I} K_i$ o es vacía o es un conjunto convexo.

Dem:
Supongamos que $\bigcap_{i \in I} K_i \neq \emptyset$. Sean $\overrightarrow{x}, \overrightarrow{y} \in \bigcap_{i \in I} K_i$, entonces $\overrightarrow{x}, \overrightarrow{y} \in K_i$, $\forall i \in I$. Por tanto, $\forall \lambda \in [0, 1]$ se verifica que $\lambda \overrightarrow{x} + (1 - \lambda)\overrightarrow{y} \in K_i, \quad \forall i \in I$ pues K_i es convexo, de donde $\lambda \overrightarrow{x} + (1 - \lambda)\overrightarrow{y} \in \bigcap_{i \in I} K_i$. Por tanto, $\bigcap_{i \in I} K_i$ es convexo. ▼

Recuérdese que un hiperplano afín en $\mathbb{R}^n$ es una variedad afín donde la dimensión del subespacio director es $n - 1$. Un hiperplano afín en $\mathbb{R}^n$ tiene por ecuación $a_1 x_1 + a_2 x_2 + \cdots + a_n x_n + b = 0$ donde $(a_1, a_2, \ldots, a_n) \neq \overrightarrow{0}$.

Definición 3.2.2
Un **semiespacio** en $\mathbb{R}^n$ es un subconjunto de la forma $a_1 x_1 + a_2 x_2 + \cdots + a_n x_n + b \leq 0$ donde $(a_1, a_2, \ldots, a_n) \neq \overrightarrow{0}$.

Ejemplo 3.2.3
El conjunto $3x + 3y \leq 3$ es un semiespacio en $\mathbb{R}^2$. ♦

Un hiperplano afín $H \equiv a_1x_1 + a_2x_2 + \cdots + a_nx_n + b = 0$ divide el espacio $\mathbb{R}^n$ en dos semiespacios $a_1x_1+a_2x_2+\cdots+a_nx_n+b \geq 0$ y $a_1x_1+a_2x_2+\cdots+a_nx_n+b \leq 0$. Si H está escrito de modo que el primer coeficiente a_i no nulo sea positivo, a dichos semiespacios se les denota por H_+ y H_-.

Teorema 3.2.3
Todo semiespacio afín de $\mathbb{R}^n$ es convexo.

Dem:
Sea un semiespacio $S \equiv a_1x_1 + a_2x_2 + \cdots + a_nx_n + b \leq 0$ y sean los puntos $(x_1, x_2, \ldots, x_n), (y_1, y_2, \ldots, y_n) \in S$, sea $\lambda \in [0, 1]$. Entonces se tiene que

$$\lambda((x_1, x_2, \ldots, x_n)+(1-\lambda)(y_1, y_2, \ldots, y_n) = (\lambda x_1+(1-\lambda)y_1, \ldots, \lambda x_n+(1-\lambda)y_n)$$

$$a_1(\lambda x_1 + (1-\lambda)y_1) + a_2(\lambda x_2 + (1-\lambda)y_2) + \cdots + a_n(\lambda x_n + (1-\lambda)y_n) + b =$$

$$= \lambda(a_1x_1+a_2x_2+\cdots+a_nx_n)+\lambda b+(1-\lambda)(a_1y_1+a_2y_2+\cdots+a_ny_n)+(1-\lambda)b =$$

$$= \lambda(a_1x_1+a_2x_2+\cdots+a_nx_n+b)+(1-\lambda)(a_1y_1+a_2y_2+\cdots+a_ny_n+b) \leq \lambda 0+(1-\lambda)0 = 0$$

pues $\lambda \geq 0$ y $1-\lambda \geq 0$. Por tanto, S es convexo. ▼

Teorema 3.2.4
Sea H un hiperplano de $\mathbb{R}^n$. Entonces H es convexo.

Dem:
El hiperplano H puede escribirse como $H = H_+ \bigcap H_-$ y puesto que H_+ y H_- son convexos, se tiene que su intersección H también lo es. ▼
Consecuencia de este último teorema es la siguiente:

Teorema 3.2.5
Toda variedad afín de $\mathbb{R}^n$ es convexa.

Dem:
Inmediata pues toda variedad afín es intersección de hiperplanos. ▼

Teorema 3.2.6
Sea $K \subset \mathbb{R}^n$ un conjunto cerrado y convexo y sea $\overrightarrow{p} \notin K$. Entonce son equivalentes:

1. $\overrightarrow{0} \notin K$.
2. $\exists \overrightarrow{u} \in K$ tal que $\langle \overrightarrow{u}, \overrightarrow{x} \rangle > 0,\ \forall \overrightarrow{x} \in K$.

Dem:
$\Leftarrow$) Si $\exists \overrightarrow{u} \in K$ tal que $\langle \overrightarrow{u}, \overrightarrow{x} \rangle > 0,\ \forall \overrightarrow{x} \in K$ evidentemente $\overrightarrow{0} \notin K$ pues $\langle \overrightarrow{u}, \overrightarrow{0} \rangle = 0$.
$\Rightarrow$)
Sea $\rho = \text{ínf}\,\{||\overrightarrow{x}||; \overrightarrow{x} \in K\}$. Considérese una sucesión de puntos $\overrightarrow{u}_i \in K$ de modo que $\text{lím}_{n\to+\infty} ||\overrightarrow{u}_n|| = \rho$. La sucesión anterior es de Cauchy, pues $\forall \overrightarrow{v}, \overrightarrow{w} \in \mathbb{R}^n$ se verifica

$$||\overrightarrow{v} + \overrightarrow{w}||^2 + ||\overrightarrow{v} - \overrightarrow{w}||^2 = 2||\overrightarrow{v}||^2 + 2||\overrightarrow{w}||^2$$

de este modo

$$||\overrightarrow{u}_i - \overrightarrow{u}_j||^2 = 2||\overrightarrow{u}_i||^2 + 2||\overrightarrow{u}_j||^2 - ||\overrightarrow{u}_i - \overrightarrow{u}_j||^2 = 2||\overrightarrow{u}_i||^2 + 2||\overrightarrow{u}_j||^2 - 4||\frac{\overrightarrow{u}_i + \overrightarrow{u}_j}{2}||^2$$

En la relación anterior se tiene que $\frac{\overrightarrow{u}_i + \overrightarrow{u}_j}{2} \in K$ y por tanto, $||\frac{\overrightarrow{u}_i + \overrightarrow{u}_j}{2}|| \geq \rho$, luego

$$||\overrightarrow{u}_i - \overrightarrow{u}_j||^2 \leq 2||\overrightarrow{u}_i||^2 + 2||\overrightarrow{u}_j||^2 - 4\rho.$$

Así, puesto que para $i, j \to +\infty$, se tiene:

$$\lim_{i,j \to +\infty} ||\overrightarrow{u}_i - \overrightarrow{u}_j||^2 \leq 2 \lim_{i \to +\infty} ||\overrightarrow{u}_i|| + 2 \lim_{j \to +\infty} ||\overrightarrow{u}_j|| - 4\rho = 4\rho - 4\rho = 0$$

por lo que la sucesión $\{\overrightarrow{u}_n\}_{n=1}^{\infty}$ es de Cauchy, y por tanto, $\exists \overrightarrow{u} \in K$ de modo que $||\overrightarrow{u}|| = \rho$. Por otra parte, se tiene que dicho punto $\overrightarrow{u}$ es único, pues sea $\overrightarrow{z} \in K$ de modo que $||\overrightarrow{z}|| = \rho$, entonces

$$||\overrightarrow{v} - \overrightarrow{w}||^2 = 2||\overrightarrow{v}||^2 + 2||\overrightarrow{w}||^2 - 4||\frac{\overrightarrow{v} - \overrightarrow{w}}{2}||^2 \leq 2||\overrightarrow{v}||^2 + 2||\overrightarrow{w}||^2 - 4\rho = 4\rho - 4\rho = 0$$

Sea $\overrightarrow{u}$ el punto de norma mínima $\rho \geq 0$ en K y sea $\overrightarrow{x} \in K$, sea $\theta \in]0, 1[$, entonces

$$0 \leq ||\theta \overrightarrow{u} + (1-\theta)\overrightarrow{x}|| - ||\overrightarrow{u}|| = \theta^2 ||\overrightarrow{u} - \overrightarrow{x}|| + 2\theta \langle \overrightarrow{u} - \overrightarrow{x}, \overrightarrow{u} \rangle,$$

y por tanto, $\forall \theta \in]0, 1[$

$$0 \leq \theta ||\overrightarrow{u} - \overrightarrow{x}|| + 2\langle \overrightarrow{u} - \overrightarrow{x}, \overrightarrow{u} \rangle,$$

de donde $\langle \overrightarrow{u}, \overrightarrow{u} \rangle - \langle \overrightarrow{x}, \overrightarrow{u} \rangle = \langle \overrightarrow{u} - \overrightarrow{x}, \overrightarrow{u} \rangle \geq 0$ y por tanto, $\langle \overrightarrow{u}, \overrightarrow{x} \rangle > \rho^2 > 0$.
▼

Teorema 3.2.7 (Primer teorema de separación)
Sea $X \subset \mathbb{R}^n$ un conjunto cerrado y convexo. Sea $\overrightarrow{p} \notin X$, entonces para algún $\overrightarrow{v} \in X$, $\overrightarrow{v} \neq \overrightarrow{0}$ se verifica $\langle \overrightarrow{v}, \overrightarrow{p} \rangle < \inf\{\langle \overrightarrow{v}, \overrightarrow{x} \rangle; \overrightarrow{x} \in X\}$.

Dem:
Sea $S = X - \overrightarrow{p}$, S es la imagen de X por una transformación afín, por tanto, S es cerrado y convexo. Puesto que $\overrightarrow{p} \notin X$, $\overrightarrow{0} \notin S$, al examinar la demostración del teorema anterior resulta que $\exists \overrightarrow{v} \in S$, $\overrightarrow{v} \neq \overrightarrow{0}$ de modo que $\langle \overrightarrow{v}, \overrightarrow{y} \rangle \geq \langle \overrightarrow{v}, \overrightarrow{v} \rangle$, $\forall \overrightarrow{y} \in S$, y como $\overrightarrow{y} = \overrightarrow{x} - \overrightarrow{p}$ se tiene que $\langle \overrightarrow{v}, \overrightarrow{x} - \overrightarrow{p} \rangle \geq \langle \overrightarrow{v}, \overrightarrow{v} \rangle$ de donde $\langle \overrightarrow{v}, \overrightarrow{v} \rangle \leq \langle \overrightarrow{v}, \overrightarrow{p} \rangle - ||\overrightarrow{v}||^2 < \inf\{\langle \overrightarrow{v}, \overrightarrow{x} \rangle; \overrightarrow{x} \in X\}$. ▼

Definición 3.2.3
Sea el espacio vectorial $\mathbb{R}^n$. Llamamos combinación lineal convexa de los vectores $\overrightarrow{x}_1, \overrightarrow{x}_2, \ldots, \overrightarrow{x}_n \in \mathbb{R}^n$ a toda expresión de la forma $\sum_{i=0}^{k} \lambda_k \overrightarrow{x}_i$ donde $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$, $\lambda_i \geq 0, \quad \forall i = 1, \ldots, k$, $\lambda_1 + \lambda_2 + \cdots + \lambda_k = 1$.

Ejemplo 3.2.4
Sean los vectores de $\mathbb{R}^2$, $(1,0)$, $(0,1)$, $(1,1)$. Determinar si los vectores $(\frac{1}{2},\frac{1}{2})$ y $(\frac{1}{3},0)$ se pueden escribir como combinación lineal convexa de los anteriores.

Para que $(\frac{1}{2},\frac{1}{2})$ sea combinación lineal convexa de los vectores $(1,0)$, $(0,1)$, $(1,1)$ deben existir escalares $\lambda_1,\lambda_2,\lambda_3\in\mathbb{R}$, $\lambda_1\geq 0$, $\lambda_2\geq 0$, $\lambda_3\geq 0$, tales que $\lambda_1+\lambda_2+\lambda_3=1$ y $\lambda_1(1,0)+\lambda_2(0,1)+\lambda_3(1,1)=(\frac{1}{2},\frac{1}{2})$. Esto es equivalente a

$$\begin{array}{rcl}\lambda_1+\lambda_2+\lambda_3 & = & 1\\ \lambda_1 \qquad +\lambda_3 & = & \frac{1}{2}\\ \lambda_1+\lambda_2 \qquad & = & \frac{1}{2}\end{array} \tag{3.1}$$

sistema que resulta compatible determinado y cuya solución viene dada por $\lambda_1=0$, $\lambda_2=\lambda_3=\frac{1}{2}$, por tanto, en el primer caso la respuesta es afirmativa.

Para que $(\frac{1}{3},0)$ sea combinación lineal convexa de los vectores $(1,0)$, $(0,1)$, $(1,1)$ deben existir escalares $\lambda_1,\lambda_2,\lambda_3\in\mathbb{R}$, $\lambda_1\geq 0$, $\lambda_2\geq 0$, tales que $\lambda_3\geq 0$ y $\lambda_1+\lambda_2+\lambda_3=1$ y $\lambda_1(1,0)+\lambda_2(0,1)+\lambda_3(1,1)=(\frac{1}{3},0)$. Esto es equivalente a

$$\begin{array}{rcl}\lambda_1+\lambda_2+\lambda_3 & = & 1\\ \lambda_1 \qquad +\lambda_3 & = & \frac{1}{3}\\ \lambda_1+\lambda_2 \qquad & = & \frac{1}{0}\end{array} \tag{3.2}$$

El sistema también es, en este caso, compatible determinado, siendo su solución $\lambda_1=-\frac{2}{3}$, $\lambda_2=\frac{2}{3}$, $\lambda_3=1$. Solución que no cumple el requisito de $\lambda_1\geq 0$ y, por tanto, la respuesta en este caso es negativa. ♦

El siguiente teorema nos ofrece una definición alternativa de conjunto convexo.

Teorema 3.2.8
Sea $K\subset\mathbb{R}^n$. Entonces K convexo si $\forall\overrightarrow{x}_1,\overrightarrow{x}_2,\ldots,\overrightarrow{x}_k\in K$ y $\forall\lambda_1,\lambda_2,\ldots,\lambda_k\in\mathbb{R}$ con $\lambda_i\geq 0$ y $\lambda_1+\lambda_2+\cdots+\lambda_k=1$ se verifica que $\lambda_1\overrightarrow{x}_1+\lambda_2\overrightarrow{x}_2+\cdots+\lambda_k\overrightarrow{x}_k\in K$.

Dem:

$\Leftarrow)$

Sean $\overrightarrow{x},\overrightarrow{y}\in K$ y sea $\alpha\in[0,1]$. Definiendo $\lambda_1=\alpha$ y $\lambda_2=1-\alpha$ se tiene que $\lambda_1\geq 0$, $\lambda_2\geq 0$ y $\lambda_1+\lambda_2=1$, por tanto, $\alpha\overrightarrow{x}+(1-\alpha)\overrightarrow{y}=\lambda_1\overrightarrow{x}+\lambda_2\overrightarrow{y}\in K$, y así K es convexo.

$\Rightarrow)$

Sean $\overrightarrow{x}_1,\overrightarrow{x}_2,\ldots,\overrightarrow{x}_k\in K$ y sea $\lambda_1,\lambda_2,\ldots,\lambda_k\in\mathbb{R}$ con $\lambda_i>0$ y $\lambda_1+\lambda_2+\cdots+\lambda_k=1$.

Si $k=1$ se tiene que $\lambda_1=1$ y, por tanto, $\lambda_1\overrightarrow{x}_1=\overrightarrow{x}_1\in K$. Supongamos que se cumple para $k-1$, entonces para k se tiene que:

$$\lambda_1\overrightarrow{x_1}+\cdots+\lambda_k\overrightarrow{x}_k=(\lambda_1+\cdots+\lambda_{k-1})\left[\frac{\lambda_1}{\lambda_1+\cdots+\lambda_{k-1}}\overrightarrow{x}_1+\right.$$
$$\left.+\cdots+\frac{\lambda_{k-1}}{\lambda_1+\cdots+\lambda_{k-1}}\overrightarrow{x}_{k-1}\right]+\lambda_k\overrightarrow{x}_k$$

Sea $\alpha_i = \frac{\lambda_i}{\lambda_1+\cdots+\lambda_{k-1}}$, $i = 1, \ldots, k-1$, entonces se tiene que $\alpha_i > 0$ y también que $\alpha_1 + \cdots + \alpha_{k-1} = \frac{\lambda_1+\cdots+\lambda_{k-1}}{\lambda_1+\cdots+\lambda_{k-1}} = 1$, de donde por hipótesis de inducción

$$\overrightarrow{y} = \frac{\lambda_1}{\lambda_1 + \cdots + \lambda_{k-1}} \overrightarrow{x}_1 + \cdots + \frac{\lambda_{k-1}}{\lambda_1 + \cdots + \lambda_{k-1}} \overrightarrow{x}_{k-1} \in K$$

así

$$\lambda_1 \overrightarrow{x_1} + \cdots + \lambda_k \overrightarrow{x}_k = (\lambda_1 + \cdots + \lambda_{k-1}) \overrightarrow{y} + \lambda_k \overrightarrow{x}_k = (1 - \lambda_k) \overrightarrow{y} + \lambda_k \overrightarrow{x}_k \in K$$

▼

Definición 3.2.4
Sea $S \subset \mathbb{R}^n$. Llamamos **envoltura convexa** de S al conjunto $\mathrm{co}(S)$ formado por las combinaciones lineales convexas de elementos de S.

Teorema 3.2.9
Sea $S \subset \mathbb{R}^n$. Entonces $\mathrm{co}(S)$ es el menor subconjunto convexo de $\mathbb{R}^n$ que contiene a S.

Dem:
Sean $\overrightarrow{y}_1, \overrightarrow{y}_2 \in \mathrm{co}(K)$. Entonces se verifica que

$$y_1 = \sum_{i=1}^{r} \lambda_i \overrightarrow{x}_i, \quad \overrightarrow{x}_i \in K, \quad \lambda_i > 0, \quad \sum_{i=1}^{k} \lambda_i = 1$$

$$y_2 = \sum_{i=1}^{s} \eta_i \overrightarrow{z}_i, \quad \overrightarrow{z}_i \in K, \quad \eta_i > 0, \quad \sum_{i=1}^{s} \eta_i = 1$$

y sea $\alpha \in [0, 1]$. Entonces

$$\alpha \overrightarrow{y}_1 + (1 - \alpha) \overrightarrow{y}_2 = \sum_{i=1}^{r} \alpha \lambda_i \overrightarrow{x}_i + \sum_{i=1}^{s} (1 - \alpha) \eta_i \overrightarrow{z}_i$$

Sea $\xi_i = \alpha \lambda_i, \quad i = 1, \ldots, r$ y sea $\xi_{r+i} = (1 - \alpha) \eta_i, \quad i = 1, \ldots, s$ y sea $\overrightarrow{v}_i = \overrightarrow{x}_i, \quad i = 1, \ldots, r$ y sea $\overrightarrow{v}_{r+i} = \overrightarrow{z}_i, \quad i = 1, \ldots, s$. Entonces

$$\alpha \overrightarrow{y}_1 + (1 - \alpha) \overrightarrow{y}_2 = \sum_{i=1}^{r+s} \xi_i \overrightarrow{v_i}$$

donde $\overrightarrow{v}_1 \in K, \forall i = 1, \ldots, r + s, \xi_i \geq 0, \forall i = 1, \ldots, r + s$ y $\sum\limits_{i=1}^{r+s} \xi_i = 1$, de donde $\alpha \overrightarrow{y}_1 + (1 - \alpha) \overrightarrow{y}_2 \in K$ y, por tanto, K es convexo.

Sea C convexo de modo que $S \in C$. Sea $\overrightarrow{y} \in \mathrm{co}(S)$. Entonces $\exists \overrightarrow{x}_1, \ldots, \overrightarrow{x}_p \in S \subset C$ y $\exists \lambda_1 \in \mathbb{R}, \quad i = 1, \ldots, p$ con $\lambda_i \geq 0$ y $\sum\limits_{r=1}^{p} \lambda_i = 1$ de modo que $\overrightarrow{y} = \sum\limits_{i=1}^{p} \lambda_i \overrightarrow{x}_i$ y como C es convexo se tiene que $\overrightarrow{y} = \sum\limits_{i=1}^{p} \lambda_i \overrightarrow{x}_i \in C$. Por tanto, $\mathrm{co}(S) \subseteq C$. ▼

Teorema 3.2.10 (Teorema de Caratheodory)
Sea $S \in \mathbb{R}^n$. Entonces si $\overrightarrow{x} \in \mathrm{co}(S)$, $\overrightarrow{x}$ puede escribirse como combinación lineal convexa de un subconjunto $\{\overrightarrow{u}_1, \ldots, \overrightarrow{u}_p\} \subset S$ donde $p \leq n+1$.

Dem:

Sea $\overrightarrow{x} \in \mathrm{co}(S)$, entonces $\exists \overrightarrow{u}_1, \ldots, \overrightarrow{u}_k$ y $\exists \lambda_1, \ldots, \lambda_k \in \mathbb{R}$ con $\lambda_i > 0$, $i = 1, \ldots, k$ y $\sum\limits_{i=1}^{k} \lambda_i = 1$ de modo que $\overrightarrow{x} = \sum\limits_{i=1}^{k} \lambda_1 \overrightarrow{u}_i$. Si $k \leq n+1$ el teorema está probado.

Si $k > n+1$, supongamos que ese k es el menor valor para el que en las condiciones anteriores $\overrightarrow{x} = \sum\limits_{i=1}^{k} \lambda_1 \overrightarrow{u}_i$. En esas condiciones se tiene: $\sum\limits_{i=1}^{k} \lambda_1 (\overrightarrow{u}_i - \overrightarrow{x}) = 0$.

Por otra parte $\exists \alpha_1, \ldots, \alpha_k$, no todos nulos, de modo que $\alpha_1 = 0$ y $\sum\limits_{i=2}^{k} \lambda_i (\overrightarrow{u}_i - \overrightarrow{x}) = 0$. Por tanto, se tiene que $\sum\limits_{i=1}^{k} (\lambda_i + t\alpha_i)(\overrightarrow{u}_i - \overrightarrow{x}) = 0$. Sea t_0 un valor de t para el cual $\varphi(t) = \mathrm{mín}\{\lambda_i + t\alpha_i; i = 2, \ldots, k\}$ es nulo. Sea $\lambda_j + t_0\alpha_j = 0$, sea $M = \sum\limits_{i=1}^{k} (\lambda_i + t_0\alpha_i)$ y sea $\beta_i = \frac{\lambda_i + t_0\alpha_i}{M}$; así, se tiene que $\alpha_i \geq 0$, con $\beta_1 > 0$, $\beta_j = 0$, $\sum\limits_{i=1}^{k} \beta_i = 1$, $\sum\limits_{i=1}^{j-1} \beta_i(\overrightarrow{u}_i - \overrightarrow{x}) + \sum\limits_{i=j+1}^{k} \beta_i(\overrightarrow{u}_i - \overrightarrow{x})$ de donde $\overrightarrow{x} = \sum\limits_{i=1}^{j-1} \beta_i \overrightarrow{u}_i + \sum\limits_{i=j+1}^{k} \beta_i \overrightarrow{u}_i$.
Por tanto, $\overrightarrow{x}$ puede ponerse como combinación lineal convexa de $k-1$ vectores de S, lo cual contradice la hipótesis de que k es mínimo. ▼

Teorema 3.2.11
Sea $S \subset \mathbb{R}^n$ compacto. Entonces $\mathrm{co}(S)$ es compacto.

Dem:

Sea S un conjunto compacto de $\mathbb{R}^n$. Sea el conjunto $V \subset \mathbb{R}^{n+1} \times \prod\limits_{i=0}^{n} S$ definido como

$$V = \{\alpha_0, \ldots, \alpha_n, \overrightarrow{u}_0, \ldots, \overrightarrow{u}_n : \alpha_i \in [0,1],\ \overrightarrow{u}_i \in S,\ i = 0, \ldots, n;\ \sum_{i=0}^{n} \alpha_i = 1\}$$

V es un subconjunto cerrado y acotado de $\mathbb{R}^{(n+1)^2}$, por tanto, V es compacto. Sea $\Psi : \mathbb{R}^{(n+1)^2} \to \mathbb{R}^n$ definida como $\Psi(\alpha_0, \alpha_1, \ldots, \alpha_n, \overrightarrow{u}_0, \overrightarrow{u}_1, \ldots, \overrightarrow{u}_n) = \sum\limits_{i=0}^{n} \alpha_i \overrightarrow{u}_i$ la cual es continua y cuya imagen es $\mathrm{co}(S)$ (teorema de Caratheodory). Puesto que V es compacto, se tiene que $\mathrm{co}(S)$ es compacto. ▼

Teorema 3.2.12
Para todo conjunto convexo $S \subset \mathbb{R}^n$ son equivalentes:

- (i) $\exists \overrightarrow{v} \in \mathbb{R}^n$ tal que $\langle \overrightarrow{v}, \overrightarrow{u} \rangle > 0\ \forall \overrightarrow{u} \in S$.
- (ii) $\overrightarrow{0} \notin \mathrm{co}(S)$.

Dem:

$\Leftarrow$) Si (ii) es falso, entonces $\exists \alpha_1, \ldots, \alpha_n \in [0,1]$ y $\exists \overrightarrow{u}_1, \ldots, \overrightarrow{u}_k \in S$ con $k \leq n+1$, y $\sum_{i=1}^{k} \alpha_i = 1$ de modo que $\overrightarrow{0} = \sum_{i=1}^{k} \overrightarrow{u}_i$. Por tanto:

$$\overrightarrow{0} = \langle \overrightarrow{v}, \overrightarrow{0} \rangle = \sum_{i=1}^{k} \langle \overrightarrow{v}, \overrightarrow{u}_i \rangle$$

y, por tanto, no todos los valores $\langle \overrightarrow{v}, \overrightarrow{u}_i \rangle$ pueden ser positivos, lo que implica que (ii) es falso.

$\Rightarrow$)

Si (ii) es cierto, se tiene por el teorema anterior que $\mathrm{co}(S)$ es compacto y por 3.2.6 se tiene que $\exists \overrightarrow{v} \in \mathbb{R}^n$ tal que $\langle \overrightarrow{v}, \overrightarrow{u} \rangle > 0 \, \forall \overrightarrow{u} \in S$. ▼

Definición 3.2.5

Sea K un conjunto convexo. Diremos que $\overrightarrow{z} \in K$ es **punto extremo** de K si $\forall \overrightarrow{x}_1, \ldots, \overrightarrow{x}_k \in K$ de modo que $\exists \lambda_1, \ldots, \lambda_k \in \mathbb{R}$ con $\lambda_i > 0$ y $\lambda_1 + \cdots + \lambda_k = 1$ de modo que la relación $\overrightarrow{z} = \sum_{i=1}^{k} \lambda_i \overrightarrow{x}_i$ implica necesariamente $\overrightarrow{x}_1 = \ldots = \overrightarrow{x}_k = \overrightarrow{z}$.

Teorema 3.2.13 (Teorema de Krein-Milman)

Todo conjunto $K \subset \mathbb{R}^n$ convexo y compacto es la clausura de la envoltura convexa de sus puntos extremos.

Dem:

Sea $n = 1$. Entonces $K = [a, b]$ y, por tanto, K es la envoltura convexa de sus puntos extremos $\{a, b\}$. Supongamos cierta la hipótesis para dimensiones menores que n, y sea $K \subset \mathbb{R}^n$. Sea $S = \{\overrightarrow{x} \in K; \overrightarrow{x}$ punto extremo de $K\}$ y sea $H = \mathrm{co}(S)$. Dado que K es convexo, se verifica que $H \subset K$. Puesto que K cerrado se tiene $\overline{H} \subset K$. Por otra parte, supongamos que $\overline{H} \neq K$, entonces $\exists \overrightarrow{p} \in K - \overline{H}$. Sin pérdida de generalidad podemos suponer que $\overrightarrow{p} = \overrightarrow{0}$ pues si no lo es, lo logramos mediante la transformación afín $A(\overrightarrow{x}) = \overrightarrow{x} - \overrightarrow{p}$. Como $\overrightarrow{0} \notin \overline{H}$, por 3.2 $\forall \overrightarrow{x} \in \overline{H}$, $\exists \overrightarrow{v} \in \mathbb{R}^n$ de modo que se verifica que $\langle \overrightarrow{x}, \overrightarrow{v} \rangle < 0$. Sea $c = \sup\{\langle \overrightarrow{x}, \overrightarrow{v} \rangle >: \overrightarrow{x} \in K\}$ de donde $c \geq 0$. Puesto que K es compacto, ese supremo se alcanza en algún punto, lo que implica que el conjunto $K' = \{\overrightarrow{x} \in K : \langle \overrightarrow{x}, \overrightarrow{v} \rangle >= c\}$ es no vacío. K' es un conjunto compacto y convexo, pues es intersección de K con el hiperplano afín $\{\overrightarrow{x} \in \mathbb{R}^n; \langle \overrightarrow{x}, \overrightarrow{v} \rangle = c\}$.

Sea $\overrightarrow{z}$ punto extremo de K'. Supongamos que $\overrightarrow{z}$ no es punto extremo de K', entonces $\exists \overrightarrow{z}_1, \overrightarrow{z}_2 \in K$ con $\overrightarrow{z} \neq \overrightarrow{z}_1$, $\overrightarrow{z} \neq \overrightarrow{z}_2$, y $\exists \lambda \in]0,1[$ de modo que $\overrightarrow{z} = \lambda \overrightarrow{z}_1 + (1-\lambda) \overrightarrow{z}_2$, a partir de lo cual se tiene que

$$c = \langle \overrightarrow{z}, \overrightarrow{v} \rangle = \langle \lambda \overrightarrow{z}_1 + (1-\lambda) \overrightarrow{z}_2, \overrightarrow{v} \rangle = \lambda \langle \overrightarrow{z}_1, \overrightarrow{v} \rangle + (1-\lambda) \langle \overrightarrow{z}_2, \overrightarrow{v} \rangle < c$$

si $\langle \overrightarrow{z}_1, \overrightarrow{v} \rangle < c$ ó $\langle \overrightarrow{z}_2, \overrightarrow{v} \rangle < c$. Por tanto, $\langle \overrightarrow{z}_1, \overrightarrow{v} \rangle = \langle \overrightarrow{z}_2, \overrightarrow{v} \rangle = c$ de donde $\overrightarrow{z}_1, \overrightarrow{z}_2 \in K'$, por tanto, $\overrightarrow{z}$ no sería punto extremo de K', lo cual es absurdo. A partir de este resultado se tiene que $\overrightarrow{z} \in S \subset \overline{H}$ y, por tanto, $c = \langle \overrightarrow{z}, \overrightarrow{v} \rangle < 0$, lo cual es absurdo. Por tanto, $K = \overline{H} = \overline{\mathrm{co}(S)}$. ▼

Teorema 3.2.14
Sea $K \subset \mathbb{R}^n$ un conjunto convexo y compacto y sea $f : \mathbb{R}^n \to \mathbb{R}$ lineal. Entonces, el máximo y el mínimo de f en K se alcanzan al menos en un punto extremo.

Dem:

Sea $M = \sup\{f(\overrightarrow{x}) : \overrightarrow{x} \in K\}$ y sea $K' = \{\overrightarrow{x} \in K : f(\overrightarrow{x}) = c\}$, puesto que K es compacto, sea $M = \text{máx}\{f(\overrightarrow{x}) : \overrightarrow{x} \in K\}$, por lo que $\exists \overrightarrow{z} \in K$ de modo que $f(\overrightarrow{z}) = M$ y, por tanto, K' es no vacío. Por otra parte, $K' \subseteq K$, por tanto, K' es acotado, y como $K' = f^{-1}(c)$ y f es continua, se tiene que K' es compacto. Por el teorema de Krein-Milman K' es combinación lineal convexa de sus puntos extremos, y por tanto, K' tiene al menos un punto extremo. Sea $\overrightarrow{x}$ punto extremo de K'. Supongamos que $\exists \overrightarrow{x}_1, \overrightarrow{x}_2 \in K$, $\overrightarrow{x}_1 \neq \overrightarrow{x}$, $\overrightarrow{x}_2 \neq \overrightarrow{x}$ de manera que $\exists \lambda \in]0,1[$ de modo que $\overrightarrow{x} = \lambda \overrightarrow{x}_1 + (1-\lambda)\overrightarrow{x}_2$, por otra parte se tiene que bien $f(\overrightarrow{x}_1) < c$, bien $f(\overrightarrow{x}_2) < c$, o ambas, pues $\overrightarrow{x}$ punto extremo en K'. De este modo, se tiene que

$$c = f(\overrightarrow{x}) = \lambda f(\overrightarrow{x}_1) + (1-\lambda) f(\overrightarrow{x}_2) < \lambda c + (1-\lambda)c = c$$

lo cual es absurdo, por tanto, $\overrightarrow{x}$ es también punto extremo en K. ▼.

3.3. Programación lineal

El problema que se plantea en programación lineal consiste en optimizar una función afín $Z = c_1x_1 + c_2x_2 + \cdots + c_nx_n + d$ sometida a un conjunto de restricciones lineales de igualdad o desigualdad dado por

$$\begin{array}{ccccccc}
a_{1,1}x_1 & + & a_{1,2}x_2 & + & \cdots & +a_{1,n}x_1 & \leq & b_1 \\
\cdots & + & \cdots & + & \cdots & +\cdots & \leq & \cdots \\
a_{p,1}x_1 & + & a_{p,2}x_2 & + & \cdots & +a_{p,n}x_1 & \leq & b_p \\
a_{p+1,1}x_1 & + & a_{p+1,2}x_2 & + & \cdots & +a_{p+1,n}x_1 & = & b_{p+1} \\
\cdots & + & \cdots & + & \cdots & +\cdots & = & \cdots \\
a_{p+q,1}x_1 & + & a_{p+q,2}x_2 & + & \ldots & +a_{p+q,n}x_1 & = & b_{p+q} \\
a_{p+q+1,1}x_1 & + & a_{p+q+1,2}x_2 & + & \cdots & +a_{p+q+1,n}x_1 & \geq & b_{p+q+1} \\
\cdots & + & \cdots & + & \cdots & +\cdots & \leq & \cdots \\
a_{p+q+r,1}x_1 & + & a_{p+q+r,2}x_2 & + & \cdots & +a_{p+q+r,n}x_1 & \geq & b_{p+q+r}
\end{array}$$

$$p, q, r, n \in \mathbb{Z}^+ \text{ con } p + q + r > 0,\ n > 0,\ c_i, d, a_{i,j}, x_i \in \mathbb{R}$$

La función Z a optimizar recibe el nombre de función objetivo, el conjunto $K \subset \mathbb{R}^n$ definido por las restricciones recibe el nombre de **conjunto factible**, y es un conjunto convexo. Las variables $x_1, x_2, \ldots, x_n$ reciben el nombre de **variables de decisión**, y los coeficientes $c_1, c_2, \ldots, c_n$, **coeficientes de costo**.

Definición 3.3.1
Sean $\overrightarrow{a}, \overrightarrow{b} \in \mathbb{R}^n$. Diremos que $\overrightarrow{a} \leq \overrightarrow{b}$ si $a_i \leq b_i$, $\forall i = 1, \ldots, n$. También diremos $\overrightarrow{a} \geq \overrightarrow{b}$ si $a_i \geq b_i$, $\forall i = 1, \ldots, n$.

Definición 3.3.2
Llamamos problema de programación lineal en **forma clásica** a un problema de la forma:

$$\begin{aligned}
&\text{maximizar } Z = \langle \overrightarrow{c}, \overrightarrow{x} \rangle + d \\
&\text{sometido a:} \\
&A\overrightarrow{x} \leq \overrightarrow{b} \\
&\overrightarrow{x} \geq \overrightarrow{0}, \quad A \in \mathbb{R}_{m\times n},\ \overrightarrow{x}, \overrightarrow{c} \in \mathbb{R}^n,\ \overrightarrow{b} \in \mathbb{R}^m,\ d \in \mathbb{R}
\end{aligned}$$

Ejemplo 3.3.1

$$\begin{aligned}
&\text{Maximizar } Z = 2x + 3y - 1 \\
&\text{sometido a:} \\
&\ \ x + 2y \leq 4 \\
&2x - \ \ y \leq 5 \\
&x, y \geq 0
\end{aligned}$$

♦

Cualquier problema de programación lineal puede ser reducido a forma clásica; para ello basta proceder según:

- Minimizar una función $Z = \langle \overrightarrow{c}, \overrightarrow{x} \rangle + d$ es equivalente a maximizar $-Z = -\langle \overrightarrow{c}, \overrightarrow{x} \rangle - d$. Además, se verifica que $Z_{\text{mín}} = -(-Z_{\text{máx}})$.
- La restricción $\sum_{j=1}^{n} a_{i,j}x_j \geq b_i$ es equivalente a la dada por $-\sum_{j=1}^{n} a_{i,j}x_j \leq -b_i$.
- La restricción $\sum_{j=1}^{n} a_{i,j}x_j = b_i$ es equivalente al par de restricciones

$$\sum_{j=1}^{n} a_{i,j}x_j \leq b_i$$

$$\sum_{j=1}^{n} a_{i,j}x_j \geq b_i$$

.

- La restricción $\left|\sum_{j=1}^{n} a_{i,j}x_j\right| \leq b_i$ es equivalente al par de restricciones

$$\sum_{j=1}^{n} a_{i,j}x_j \leq b_i$$

$$\sum_{j=1}^{n} a_{i,j}x_j \geq -b_i$$

.

- Si x_i es una variable de decisión que no requiere ser mayor o igual que cero, dicha variable puede ser reemplazada por la diferencia de dos variables positivas $u_i,\ v_i \geq 0$, esto es $x_i = u_i - v_i$ con $u_i,\ v_i \geq 0$.

Ejemplo 3.3.2
Escribir en su forma clásica el problema

$$\begin{aligned}
&\text{Minimizar } Z = 2x + 3y - 1\\
&\text{sometido a:}\\
&\ \ x + 2y\ \ = 1\\
&|2x - \ \ y| \geq 5\\
&x, y \in \mathbb{R}
\end{aligned}$$

Aplicando el procedimiento anterior se obtiene el problema en su **forma clásica** como:

$$\begin{aligned}
&\text{Sea}\\
&x = x_1 - x_2\\
&y = x_3 - x_4\\
&\text{donde } x_1, x_2, x_3, x_4 \geq 0
\end{aligned}$$

$$\begin{aligned}
&\text{Maximizar } Z_1 = -2x + 2x_2 - 3x_3 + 3x_4 + 1\\
&\text{sometido a:}\\
&x_1 - x_2 + 2x_3 - 2x_4 \leq 1\\
&x_1 - x_2 + 2x_3 - 2x_4 \geq 1\\
&2x_1 - 2x_2 - x_3 + x_4 \leq 5\\
&2x_1 - 2x_2 - x_3 + x_4 \geq -5\\
&\text{donde } x_1, x_2, x_3, x_4 \geq 0
\end{aligned}$$

♦

Definición 3.3.3
Llamamos problema de programación lineal en la **forma estándar** a un problema de la forma:

$$\begin{aligned}
&\text{Maximizar la función } Z = \langle \overrightarrow{c}, \overrightarrow{x} \rangle + d\\
&\text{sometido a las restricciones:}\\
&A\overrightarrow{x} = \overrightarrow{b}\\
&\overrightarrow{x} \geq \overrightarrow{0},\ \overrightarrow{b} \geq \overrightarrow{0}\\
&A \in \mathbb{R}_{n,m},\ \overrightarrow{x} \in \mathbb{R}^m,\ \overrightarrow{b} \in \mathbb{R}^n,\ \overrightarrow{c} \in \mathbb{R}^m, d \in \mathbb{R}
\end{aligned}$$

Todo problema de programación lineal puede ser llevado a su forma estándar a partir de su formulación clásica teniendo en cuenta las siguientes consideraciones

- Toda desigualdad puede ser transformada en igualdad mediante la adición o sustracción de una variable positiva llamada **variable de holgura**.
- Toda desigualdad puede ser cambiada de sentido sin mas que multiplicar ambos miembros por -1.

Ejemplo 3.3.3
Sea el problema dado en su forma clásica como:

$$\begin{aligned}
&\text{Maximizar } Z = -2x + 2x_2 - 3x_3 + 3x_4 + 1\\
&\text{sometido a:}\\
&x_1 - x_2 + 2x_3 - 2x_4 \leq 1\\
&x_1 - x_2 + 2x_3 - 2x_4 \geq 1\\
&2x_1 - 2x_2 - x_3 + x_4 \leq 5\\
&2x_1 - 2x_2 - x_3 + x_4 \geq -5\\
&\text{donde } x_1, x_2, x_3, x_4 \geq 0
\end{aligned}$$

aplicando las transformaciones anteriores se tiene el problema en su forma estándar como:

$$\begin{aligned}
&\text{Maximizar } Z = -2x + 2x_2 - 3x_3 + 3x_4 + 1\\
&\text{sometido a:}\\
&\quad x_1 - x_2 + 2x_3 - 2x_4 + x_5 = 1\\
&\quad x_1 - x_2 + 2x_3 - 2x_4 - x_6 = 1\\
&\quad 2x_1 - 2x_2 - x_3 + x_4 + x_7 = 5\\
&-2x_1 + 2x_2 + x_3 - x_4 + x_8 = 5\\
&\text{donde } x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8 \geq 0
\end{aligned}$$

♦

Definición 3.3.4
Sea un problema de programación lineal dado en su forma estándar y sea $K \subset \mathbb{R}^n$ el conjunto dado por las restricciones del problema

$$\left\{\overrightarrow{x} \in \mathbb{R}^n \,:\, A\overrightarrow{x} = \overrightarrow{b},\ \overrightarrow{x} \geq \overrightarrow{0}\right\}$$

Diremos que un punto $\overrightarrow{x}$ es **punto básico** del problema si $\exists\sigma \in S_n$ de modo que la matriz $A \in \mathbb{R}_{n\times m}$ puede descomponerse como $A = [D|E]$ donde $D = \{d_{i,\sigma j}\}$, $i = 1, \ldots, m; j = 1, \ldots, m$, es matriz regular, $E = \{e_{i,\sigma j}\}$, $i = 1, \ldots, m; j = m + 1, \ldots, n$, de modo que $\overrightarrow{x}$ satisface $x_{\sigma_i} = 0$, $\forall i = m + 1, \ldots, n$ y $D\overrightarrow{y} = \overrightarrow{b}$ donde $\overrightarrow{i} = (x_{\sigma_1}, \ldots, x_{\sigma_m})^t$. Si además $\overrightarrow{x} \in K$ diremos que $\overrightarrow{x}$ es un **punto factible básico**.

El nombre de básico se debe a que las columnas $\sigma_1, \sigma_2, \ldots, \sigma_m$ de la matriz forman una base de $\mathbb{R}^m$.

Teorema 3.3.1
Un punto básico $\overrightarrow{x}$ es factible si, y sólo si, $x_i \geq 0, \forall i = 1, \ldots, n$.

Dem:
En primer lugar, si $\overrightarrow{x}$ es básico se tiene que

$$A\overrightarrow{x} = D \begin{bmatrix} x_{\sigma_1} \\ \ldots \\ x_{\sigma_m} \end{bmatrix} + E \begin{bmatrix} x_{\sigma_{m+1}} \\ \ldots \\ x_{\sigma_n} \end{bmatrix} = \overrightarrow{b} + \overrightarrow{0} = \overrightarrow{b}$$

y, por tanto, $\overrightarrow{x} \in K$ si, y sólo si, $x_i \geq 0, \forall i = 1, \ldots, n$. ▼

Definición 3.3.5
Diremos que un **punto factible** es **no degenerado** si el cardinal de $I(\overrightarrow{x})$ es n, esto es, $\overrightarrow{x}$ tiene exactamente n componentes mayores que cero. En caso contrario diremos que $\overrightarrow{x}$ es degenerado.

Definición 3.3.6
Sea un problema de programación lineal dado en su forma estándar y sea $K \subset \mathbb{R}^n$ el conjunto dado por las restricciones del problema

$$\left\{ \overrightarrow{x} \in \mathbb{R}^n \, : \, A\overrightarrow{x} = \overrightarrow{b}, \, \overrightarrow{x} \geq \overrightarrow{0} \right\}$$

y sea $\overrightarrow{x} \in K$ un punto factible. Llamamos $I(\overrightarrow{x}) = \{i \in \{1, \ldots, n\} \, : \, x_i > 0\}$.

Teorema 3.3.2 (Teorema fundamental de la programación lineal)
Sea un problema de programación lineal dado en su forma estándar y sea $\overrightarrow{A}_i$ el vector dado por la i-ésima columna de la matriz A. Sea $\overrightarrow{x} \in K$ un punto factible. Entonces las siguientes afirmaciones son equivalentes:

1. $\overrightarrow{x}$ es un punto extremo de K.
2. $\{\overrightarrow{A}_i \, : \, i \in I(\overrightarrow{x})\}$ es un conjunto linealmente independiente.

Dem:
$2) \Rightarrow 1)$
Supongamos que $\overrightarrow{x}$ no es punto extremo, entonces $\exists \overrightarrow{u}, \overrightarrow{v} \in K - \{\overrightarrow{x}\}$ y $\exists \lambda \in]0, 1[$ de modo que $\overrightarrow{x} = \lambda \overrightarrow{u} + (1 - \lambda)\overrightarrow{v}$.
Para cada $i \notin I(\overrightarrow{x})$ se tiene que $x_i = 0$, por tanto, $0 = \lambda u_i + (1 - \lambda)v_i$ y dado que $\overrightarrow{u}, \overrightarrow{v} \in K$ se tiene que $\overrightarrow{u} \geq \overrightarrow{0}$, $\overrightarrow{v} \geq \overrightarrow{0}$, por tanto, es necesario que $u_i = v_i = 0$.
Por otra parte, se verifica que $A\overrightarrow{u} = \overrightarrow{b}$, y $A\overrightarrow{u} = \overrightarrow{b}$ por tanto, $A\overrightarrow{u} - A\overrightarrow{v} = \overrightarrow{0}$ lo que es equivalente a $\sum_{i=0}^{n} (u_i - v_i)\overrightarrow{A}_i = \sum_{i \in I(\overrightarrow{x})} (u_i - v_i)\overrightarrow{A}_i = \overrightarrow{0}$.
Como $\{\overrightarrow{A}_i \, : \, i \in I(\overrightarrow{x})\}$ es linealmente independiente, es necesario que se verifique que $u_i - v_i = 0, \forall I(\overrightarrow{x})$ y por tanto, $u_i = v_i, \forall I(\overrightarrow{x})$, de donde $\overrightarrow{u} = \overrightarrow{v}$; con lo cual $\overrightarrow{x} = \lambda \overrightarrow{u} + (1 - \lambda)\overrightarrow{v} = \lambda \overrightarrow{u} + (1 - \lambda)\overrightarrow{u} = \overrightarrow{u}$, pero $\exists \overrightarrow{u}, \overrightarrow{v} \in K - \{\overrightarrow{x}\}$, lo cual es una contradicción. Por tanto, $\overrightarrow{x}$ es punto extremo de K.
$1) \Rightarrow 2)$

Supóngase que el conjunto $\{\overrightarrow{A}_i \ : \ i \in I(\overrightarrow{x})\}$ es no libre, entonces $\exists w_i \in \mathbb{R}$, $i \in I(\overrightarrow{x})$ no todos nulos de modo que $\sum\limits_{I(\overrightarrow{x})} w_i \overrightarrow{A}_i = 0$, y sea $w_i = 0$, $\forall i \notin I(\overrightarrow{x})$. Denominando como $\overrightarrow{w}$ al vector cuyas componentes viene dadas por w_i, $i = 1, \ldots, n$ se tiene que $\forall \lambda \in \mathbb{R}^+$, por tanto, se verifica que $A(\overrightarrow{x} \pm \lambda \overrightarrow{w}) = A\overrightarrow{x} \pm \lambda \overrightarrow{w} = A\overrightarrow{x} = \overrightarrow{b}$. Sea $\alpha < \text{máx}\{\lambda \in \mathbb{R}^+ \ : \ x_i \pm \lambda w_i = 0, \forall i \in I(\overrightarrow{x})\}$. Entonces, se verifica que $\overrightarrow{u} = \overrightarrow{x} + \alpha \overrightarrow{w} \in K$ y $\overrightarrow{v} = \overrightarrow{v} \alpha \overrightarrow{w} \in K$ y, por tanto, $\frac{1}{2}\overrightarrow{u} + \frac{1}{2}\overrightarrow{v} = \overrightarrow{x}$ lo cual no es posible pues $\overrightarrow{x}$ es un punto extremo de k, por tanto, $\{\overrightarrow{A}_i \ : \ i \in I(\overrightarrow{x})\}$ es libre. ▼

Definición 3.3.7
Sea un problema de programación lineal escrito en forma estándar y sea $\{\overrightarrow{A}_i \ : \ i \in I(\overrightarrow{x})\}$ base de $\mathbb{R}^n$. Entonces, llamamos **solución básica** asociada a la solución del sistema que satisface $x_i = 0 \, \forall \, i \notin I(\overrightarrow{x})\}$. A las variables x_i de modo que $i \in I(\overrightarrow{x})$ se les **denomina variables básicas**, siendo denominadas las restantes x_i como **variables no básicas**.

3.3.1. Algoritmo del simplex

La resolución de un problema de programación lineal puede abordarse por diversos procedimientos. La solución del problema de programación lineal, si ésta existe, debe alcanzarse en un punto extremo. En este curso se abordará el llamado algoritmo del simplex (Dantzing, 1948), el cual consta de dos etapas:

- Encontrar un punto extremo de K.
- Cambiar de punto extremo de K de modo que se incremente el valor de la función objetivo Z.

Para la resolución del algoritmo del simplex debemos tener en cuenta el siguiente teorema.

Teorema 3.3.3
Sea un problema de programación lineal dado en su forma estándar. Sea la función objetivo $Z = \sum\limits_{i=0}^{n} c_i x_i$ escrita de modo que si x_i es variable básica, $c_i = 0$ siendo $c_i \leq 0$ en otro caso. Entonces, Z alcanza su máximo en K en el punto $\overrightarrow{x}$ extremo, tal que sus coordenadas no básicas se anulan.

Dem:
Dado que $\forall \overrightarrow{x} \in K$ se verifica que $x_i \geq 0$, $i = 1, \ldots, n$, entonces

$$\forall \overrightarrow{y} \in K, \, Z(\overrightarrow{y}) = \sum_{i=1}^{n} c_i y_i = \sum_{i \in I(\overrightarrow{x})} c_i y_i \leq 0$$

Por otra parte, el valor cero es alcanzado pues

$$Z(\overrightarrow{x}) = \sum_{i=1}^{n} c_i x_i = \sum_{i \in I(\overrightarrow{x})} c_i x_i = 0$$

por tanto, dicho punto básico es óptimo. ▼

Para encontrar un punto extremo se aplica el teorema fundamental de la programación lineal, según el cual un punto es extremo si, y sólo si, es punto factible básico, para lo cual bastará que el conjunto de columnas de A correspondientes a los índices de $I(\overrightarrow{x})$ sea linealmente independiente, y que se satisfaga $x_i \geq, 0 \, \forall i = 1, \dots, n$. En este apartado supondremos no degeneración, así las columnas anteriores formarán una base de $\mathbb{R}^m$. En caso de degeneración, ésta se puede evitar sumando al segundo miembro de cada ecuación una cantidad ϵ^i donde i representa la fila correspondiente a la ecuación y ϵ una cantidad positiva suficientemente pequeña.

La forma habitual de dar las restricciones de un problema de programación lineal es mediante un conjunto de desigualdades lineales para las variables de decisión. Estas restricciones pueden ser transformadas en un conjunto de desigualdades lineales para un conjunto de variables positivas. Para llevar dichas desigualdades a la formulación estándar, las escribiremos, en primer lugar, dejando en el primer miembro la parte correspondiente a las variables y en el segundo miembro los términos independientes, de modo que estos resulten mayores o iguales que cero. Una vez escritas las restricciones de esta manera, se procede a introducir un conjunto de variables positivas, llamadas, como es sabido, variables de holgura, las cuales mediante su adición o sustracción permiten transformar las desigualdades en igualdades. Si las desigualdades escritas en la forma indicada son todas de la forma $\leq$, entonces habrá una variable de holgura por cada ecuación, de modo que si escribimos como es habitual las variables de holgura al final de la matriz de la forma estándar, las m últimas columnas resultan ser

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

Por tanto, el punto resultante de igualar a cero las variables correspondientes a las columnas $1, 2, \dots, n - m$ será un punto factible básico; dicho punto resulta $\overrightarrow{x} = (0, 0, \dots, 0, b_1, \dots, b_m)^t$.

En el caso de que hayan aparecido desigualdades de tipo $\geq$ el proceso anterior ya no será posible, pues las columnas correspondientes a las variables de holgura incluirán vectores de la forma $(0, \dots, 0, -1, 0, \dots, 0)^t$, por lo que la solución básica consistente en anular todas las variables excepto las de holgura no será factible por ser negativos los valores correspondientes a las columnas procedentes de las desigualdades de tipo $\leq$.

Uno de los métodos existentes para la obtención de un punto factible básico inicial consiste en la introducción de un conjunto de variables positivas llamadas variables artificiales. El procedimiento que se sigue es el siguiente: en cada ecuación en la que la variable de holgura aparece con signo menos, se añade una nueva variable artificial. De este modo, el conjunto de columnas correspondientes a las variables de holgura con signo positivo y a las variables artificiales formará una base de $\mathbb{R}^n$.

El problema inicial tendrá puntos extremos si y sólo si existen puntos extremos en el nuevo conjunto de restricciones de modo que para cada variable artificial se tenga $x_i = 0$. El siguiente ejemplo ilustra esta técnica.

Ejemplo 3.3.4
Sea un problema de programación lineal donde el conjunto $K \subset \mathbb{R}^n$ de restricciones está definido mediante las desigualdades

$$\begin{aligned} 3x + 2y + z &\geq 1 \\ 2x + 2y \quad\; &\leq 20 \\ z \quad\;\; &\geq 1 \\ x, y &\geq 0 \end{aligned}$$

En primer lugar, se escribe el problema en su formulación estándar haciendo $x = x_1$, $y = x_2$, $z = x_3$ (x, y, x son no negativas) e introduciendo las variables de holgura $x_4, x_5, x_6 \geq 0$ de modo que

$$\begin{aligned} 3x_1 + 2x_2 + x_3 - x_4 \qquad\quad &= \;\,1 \\ 2x_1 + 2x_2 \qquad\qquad + x_5 \quad\; &= 20 \\ x_3 \qquad\qquad\qquad\quad - x_6 &= \;\,1 \\ x_1, x_2, x_3, x_4, x_5, x_6 &\geq 0 \end{aligned}$$

lo que en forma matricial resulta

$$\begin{bmatrix} 3 & 2 & 1 & -1 & 0 & 0 \\ 2 & 2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} 1 \\ 20 \\ 1 \end{bmatrix}, \quad x_1, x_2, x_3, x_4, x_5, x_6 \geq 0$$

obsérvese que si se hace $x_1 = x_2 = x_3 = 0$ resulta $x_4 = -1$, $x_5 = 20$, $x_6 = -1$ por tanto, el punto no es factible.

Para lograr un punto factible inicial se introducen las variables artificiales $x_7, x_8 \geq 0$ de donde resulta el conjunto $K' \subset \mathbb{R}$ definido por las ecuaciones

$$\begin{bmatrix} 3 & 2 & 1 & -1 & 0 & 0 & 1 & 0 \\ 2 & 2 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{bmatrix} = \begin{bmatrix} 1 \\ 20 \\ 1 \end{bmatrix}, \quad x_1, x_2, x_3, x_4, x_5, x_6 \geq 0$$

donde $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8 \geq 0$, lo que define un conjunto $K' \subset \mathbb{R}^8$.

Resulta inmediato que el conjunto K es la proyección sobre $\mathbb{R}^6$ del conjunto $\{\overrightarrow{x} \in K' : x_7 = x_8 = 0\}$. Una forma de encontrar tales puntos es mediante la resolución del siguiente problema auxiliar de programación lineal.

Maximizar la función $Z_{aux} = -\sum \text{variables artificiales} = -x_7 - x_8$

sometida a las restricciones

$$\begin{bmatrix} 3 & 2 & 1 & -1 & 0 & 0 & 1 & 0 \\ 2 & 2 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{bmatrix} = \begin{bmatrix} 1 \\ 20 \\ 1 \end{bmatrix}$$

$\overrightarrow{x} \in \mathbb{R}^8,\ x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8 \geq 0$

Evidentemente, K será no vacío si y sólo si la función auxiliar Z_{aux} alcanza como máximo el valor cero en un punto en que $x_7 = x_8 = 0$. ♦

Una vez resuelto este problema, del que sí se tiene un punto factible básico inicial cuyas variables básicas vienen dadas por las columnas correspondientes a las variables de holgura que aparecen con signo $+$ y las correspondientes a las variables artificiales, se dispondrá de un punto factible básico inicial. Con ello, como se verá más adelante, el problema se reduce a la segunda etapa.

3.3.2. Algoritmo del simplex: forma abstracta

Considérese un problema de programación lineal dado en su forma estándar como $\max Z = \langle \overrightarrow{c}, \overrightarrow{x} \rangle + d$ de modo que $A\overrightarrow{x} = \overrightarrow{b}$, $A \in \mathbb{R}_{m\times n}$, $\overrightarrow{x}, \overrightarrow{c} \in \mathbb{R}^n$, $\overrightarrow{b} \in \mathbb{R}^m$, $\overrightarrow{x} \geq \overrightarrow{0}$, $\overrightarrow{b} \geq \overrightarrow{0}$.

Sea $\overrightarrow{x} \in K$ punto extremo no degenerado. Por 3.3 se tiene que el conjunto de vectores $\{\overrightarrow{A}_i\ i \in I(\overrightarrow{x})\}$ es linealmente independiente; además, por la hipótesis de no degeneración es base. Entonces para cada $j = 1, \ldots, m$ existen escalares $D_{i,j}$ tales que:

$$\overrightarrow{A}_j = \sum_{i \in I(\overrightarrow{x})} D_{i,j} \overrightarrow{A}_i \quad \forall j = 1, \ldots, m.$$

Definiendo $D_{i,j} = 0$ si $i \notin I(\overrightarrow{x})$ se puede escribir $\forall j = 1, \ldots, m$

$$\overrightarrow{A}_j = \sum_{i=1}^{n} D_{i,j} \overrightarrow{A}_i \quad \text{o } A = A \cdot D^T.$$

Sea $\overrightarrow{f} = D^t \overrightarrow{c}$. Entonces $\forall q \notin I(\overrightarrow{x})$ y $\forall \lambda \in \mathbb{R}$ se verifica que $\overrightarrow{A}_q = \sum_{i=1}^{n} D_{j,i} A_i$, y

por tanto

$$\overrightarrow{b} = A\overrightarrow{x} + \lambda\left(\overrightarrow{A}_q - \sum_{i=1}^{n} D_{j,i}A_i\right) = \sum_{i=1,n} x_i\overrightarrow{A}_i + \lambda\sum_{i=1}^{n}\delta_{i,q}\overrightarrow{A}_i - \lambda\sum_{i=1}^{n} D_{j,i}A_i =$$

$$= \sum_{i=1}^{n}\left(x_i + \lambda\sum_{i=1}^{n}\delta_{i,q} - \lambda\sum_{i=1}^{n} D_{j,i}\right)A_i = \sum_{i=1}^{n} y_i\overrightarrow{A}_i = A\overrightarrow{y}$$

donde $\overrightarrow{y} = \overrightarrow{x} - \lambda\overrightarrow{D}_q + \lambda\delta_q$ y $\delta_{i,j}$ es la delta de Kronocher, $D_q = (D_{1,q}, \ldots, D_{n,q})^t$ y $\delta_q = (\delta_{1,q}, \ldots, \delta_{n,q})^t$.

Puesto que $q \notin I(\overrightarrow{x})$, se verifica que $D_{q,q} = 0$ y $x_q = 0$, entonces $y_q = x_q - \lambda D_{q,q} + \lambda\delta_q = \lambda$. Para que $y \in K$ es necesario, por una parte, que $\lambda \geq 0$. Por otra parte, se tiene que

$$\langle\overrightarrow{c}, \overrightarrow{y}\rangle = \sum_{i=1}^{n} c_i x_i - \lambda\sum_{i=1}^{n}\langle\overrightarrow{c}, \overrightarrow{D}_q\rangle + \lambda\sum_{i=1}^{n} c_i\delta_{i,q}$$

de donde

$$\langle\overrightarrow{c}, \overrightarrow{y}\rangle = \langle\overrightarrow{c}, \overrightarrow{x}\rangle - \lambda\langle\overrightarrow{c}, \overrightarrow{D}_q\rangle + \lambda c_q = \langle\overrightarrow{c}, \overrightarrow{x}\rangle + \lambda(c_q - f_q).$$

Para incrementar el valor de la función objetivo, el valor de q debe tomarse de modo que $c_q - f_q \geq 0$. Nótese que si $\overrightarrow{c} \leq \overrightarrow{f}$ no es posible incrementar el valor de la función objetivo siendo en ese caso $\overrightarrow{x}$ una solución óptima. Cuando no se cumple la condición $\overrightarrow{c} \leq \overrightarrow{f}$, se toma q de modo que $c_q - f_q = \text{máx}\{c_i - f_i \ : \ i = 1, \ldots, n\}$, eligiéndose uno de ellos si hay varios.

Para encontrar, siguiendo este método, un nuevo punto extremo, debe tenerse en cuenta que el nuevo punto $\overrightarrow{y}$ debe ser tal que $I(\overrightarrow{y})$ tenga a lo sumo m elementos. Como $\overrightarrow{y} \subset \overrightarrow{x} \cap \{q\}$, bastará con encontrar un $\lambda \geq 0$ de modo que exista un valor $i \in \overrightarrow{x}$ tal que $x_i - \lambda D_{i,q} = 0$ y $x_j - \lambda D_{j,q} \leq 0$ si $j \neq i$. En caso de ser $D_{i,q} \leq 0$ $\forall i \in I(\overrightarrow{x})$ se verifica que $\forall\lambda \geq 0$, $x_i - \lambda D_{i,q} > 0$, por tanto, $\overrightarrow{y} \in K$ y dado que $Z(\overrightarrow{y}) = \langle\overrightarrow{c}, \overrightarrow{x}\rangle + \lambda(c_q - f_q) + d$ se tiene que $\lim\limits_{\lambda\to+\infty} Z(\overrightarrow{y}) = +\infty$ y, por tanto, se trata de un problema no acotado.

Si $\{i \in I(\overrightarrow{x}) \ : \ D_{i,q} > 0\} \neq \emptyset$ sea $\lambda = \text{mín}\left\{\frac{x_i}{D_{i,q}} \ : \ D_{i,q} > 0\right\}$, entonces se tiene que $y_i \leq 0$ $\forall i \in I(\overrightarrow{x}) \cap \{q\}$, por lo que $\overrightarrow{y} \in K$.

Sea p un valor para el que se alcanza dicho mínimo, entonces se verifica $I(\overrightarrow{y}) \subset (I(\overrightarrow{x}) \cap \{q\}) - \{p\}$. Para demostrar que $\overrightarrow{y}$ es punto extremo de K bastará demostrar que el conjunto de columnas $\{\overrightarrow{A}_i \ : \ i \in I(\overrightarrow{y})\}$ es linealmente independiente; para ello, sea $\sum\limits_{i\in I(\overrightarrow{y})} \beta_i\overrightarrow{A}_i = 0$, entonces haciendo $\beta_p = 0$ se verifica

$$\sum_{i\in I(\overrightarrow{y})} \beta_i\overrightarrow{A}_i = \sum_{i\in I(\overrightarrow{x})} \beta_i\overrightarrow{A}_i + \beta_q\overrightarrow{A}_q = 0$$

Si $\beta_q = 0$ se tiene que $\beta_i = 0$ $\forall i \in I(\overrightarrow{x})$ pues $\overrightarrow{x}$ es punto extremo y, por tanto, las columnas correspondientes de la matriz A son linealmente independientes. Si

$\beta_q \neq 0$, se verifica que $\overrightarrow{A}_q = \sum\limits_{i \in I(\overrightarrow{x})} \frac{\beta_i}{\beta_q} \overrightarrow{A}_i$, pero por otra parte también se tiene que $\overrightarrow{A}_q = \sum\limits_{i \in I(\overrightarrow{x})} D_{i,q} \overrightarrow{A}_i$ y dado que los vectores del segundo miembro son linealmente independientes, los coeficientes de la combinación lineal deben ser únicos. Así $\frac{\beta_i}{\beta_q} = D_{i,q}\ \forall i \in I(\overrightarrow{x})$, por tanto, debe verificarse que $\frac{\beta_p}{\beta_q} = D_{p,q}$ lo cual no es posible pues $\beta_p = 0$ y $D_{p,q} > 0$, por tanto, $\beta_q = 0$.

3.3.3. Algoritmo del simplex: método tabular

A continuación se expone un método tabular para resolver problemas de programación lineal mediante el método del simplex.

Sea un problema de programación lineal dado en su forma estándar $\max Z = \langle \overrightarrow{c}, \overrightarrow{x} \rangle + d$ sometido a las restricciones $A\overrightarrow{x} = \overrightarrow{b}$ donde $\overrightarrow{x}, \overrightarrow{x} \in \mathbb{R}^n$, $A \in \mathbb{R}_{m \times n}$, $\overrightarrow{b} \in \mathbb{R}^m$, $d \in \mathbb{R}$, $\overrightarrow{x} \geq \overrightarrow{0}$, $\overrightarrow{b} \geq \overrightarrow{0}$. El método tabular consiste en escribir el problema de modo que:

- En la primera línea se escribe los coeficientes de la función $-Z + \langle \overrightarrow{c}, \overrightarrow{x} \rangle = -d$.
- Las siguientes líneas deben contener cero en la primera columna (la correspondiente Z) y a continuación los coeficientes de las ecuaciones $A\overrightarrow{b} = \overrightarrow{b}$.
- Cada línea (de la dos en adelante) contiene una, y sólo una, variable básica. Cada línea contiene una variable básica distinta.
- La última columna corresponde a los términos independientes de las restricciones, los cuales deben ser mayores o iguales que cero.
- La función objetivo deberá expresarse en función de variables no básicas.

Ejemplo 3.3.5
Sea el problema de programación lineal $\max Z = 2x + 3y + z + 3$ sometido a las restricciones

$$
\begin{aligned}
&x + y + z \leq 7 \\
&x + y + z \geq -2 \\
&x + 2y \leq 5 \\
&x \leq 4 \\
&x, y \geq 0
\end{aligned}
$$

Para reducir el problema a variables positivas, sea $x = x_1$, $y = x_2$, $z = x_3 - x_4$ con $x_1, x_2, x_3, x_4 \geq 0$.

Para reducir el problema a su forma estándar, las restricciones deberán ser escritas de modo que $\overrightarrow{b} \geq 0$; así se tiene

$$\begin{aligned}
&\text{máx}\, Z = 2x_1 + 3x_2 + x_3 - x_4 + 3 \\
&\text{sometido a :} \\
&x_1 + x_2 + x_3 - x_4 \leq 7 \\
&- x_1 - x_2 - x_3 + x_4 \leq 2 \\
&x_1 + 2x_2 \leq 5 \\
&x_1 \leq 4 \\
&x_1, x_2, x_3, x_4 \geq 0
\end{aligned}$$

Introduciendo las variables de holgura $x_5, x_6, x_7, x_8 \geq 0$ se tiene el problema en su forma estándar como

$$\begin{aligned}
&\text{máx}\, Z = 2x_1 + 3x_2 + x_3 - x_4 + 3 \\
&\text{sometido a :} \\
&x_1 + x_2 + x_3 - x_4 + x_5 = 7 \\
&- x_1 - x_2 - x_3 + x_4 + x_6 = 2 \\
&x_1 + 2x_2 + x_7 = 5 \\
&x_1 + x_8 = 4 \\
&x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8 \geq 0
\end{aligned}$$

Puesto que todas las variables de holgura aparecen con signo $+$ podemos tomar éstas como variables básicas iniciales, con lo cual el problema se puede expresar de modo tabular como

	-1	2	3	1	-1	0	0	0	0	-3
x_5	0	1	1	1	-1	1	0	0	0	7
x_6	0	-1	-1	-1	1	0	1	0	0	2
x_7	0	1	2	0	0	0	0	1	0	5
x_8	0	1	0	0	0	0	0	0	1	4

La primera columna indica la variable básica correspondiente a cada fila, la segunda corresponde a la variable Z, las columnas de la tres hasta la diez corresponden a las variables $x_1, \ldots, x_8$ y la última, a los términos independientes. En muchas ocasiones, por simplicidad, se prescinde de las dos primeras columnas (de una e incluso de las dos).

Para aplicar el algoritmo del simplex directamente sobre la tabla debe procederse del siguiente modo:

1. Si todos los coeficientes de la función objetivo son menores o iguales que cero, el punto básico actual es óptimo, con lo que finaliza el algoritmo.

2. Elíjase la variable básica x_j que aparezca con mayor coeficiente en la función objetivo (si hay varias tomar una de ellas). Esta variable se tomará como nueva variable básica.

3. Si el coeficiente $a_{i,j}$ correspondiente a la i-ésima ecuación es mayor que cero, dividir toda la fila por $a_{i,j}$. Si no existe ningún coeficiente $a_{i,j} > 0$ el problema es no acotado finalizando el algoritmo.

4. De las filas en que $a_{i,j} = 1$ elíjase la k-ésima de modo que $b_k = \text{mín}\{b_i | \, a_{i,j} = 1, \, i = 1, \ldots, m\}$ (si hay varias que cumplen la condición escójase una de ellas). La variable que saldrá de la base será la variable básica correspondiente a la k- ésima fila.

5. En toda la columna j hacer el elemento $a_{i,j} = 0$ si $i \neq k$ utilizando para ello la fila k, y el proceso de eliminación gaussiana. Hacer también cero, por idéntico procedimiento, el coeficiente c_j de la función objetivo.

Ejemplo 3.3.6

Considérese el problema de programación lineal dado por la tabla anterior, y obténgase una solución óptima.

Dada la tabla (donde no se ha escrito por simplicidad la columna correspondiente a Z) se observa que no todos los coeficientes de la función objetivo son menores o iguales a cero

	2	3	1	-1	0	0	0	0	-3
x_5	1	1	1	-1	1	0	0	0	7
x_6	-1	-1	-1	1	0	1	0	0	2
x_7	1	2	0	0	0	0	1	0	5
x_8	1	0	0	0	0	0	0	1	4

Se observa que el mayor coeficiente positivo de la función objetivo es el correspondiente a la tercera columna; así, deberá entrar x_2 como nueva variable básica. Haciendo uno los elementos positivos de la tercera columna correspondientes a las ecuaciones se tiene

	2	3	1	-1	0	0	0	0	-3
x_5	1	1	1	-1	1	0	0	0	7
x_6	-1	-1	-1	1	0	1	0	0	2
x_7	$\frac{1}{2}$	1	0	0	0	0	$\frac{1}{2}$	0	$\frac{5}{2}$
x_8	1	0	0	0	0	0	0	1	4

El menor término independiente de las ecuaciones en el que el coeficiente $a_{i,2}$ es la unidad es el correspondiente a la tercera ecuación, por tanto, deberá salir de la base la variable x_7; así se tiene

	$\frac{1}{2}$	0	1	-1	0	0	$-\frac{3}{2}$	0	$-\frac{21}{2}$
x_5	$\frac{1}{2}$	0	1	-1	1	0	$-\frac{1}{2}$	0	$\frac{9}{2}$
x_6	$\frac{3}{2}$	0	-1	1	0	1	$\frac{1}{2}$	0	$\frac{9}{2}$
x_2	$\frac{1}{2}$	1	0	0	0	0	$\frac{1}{2}$	0	$\frac{5}{2}$
x_8	1	0	0	0	0	0	0	1	4

Los coeficientes de la función objetivo no son en su totalidad menores o iguales a cero, por tanto, se toma el de mayor valor que es el correspondiente a la tercera columna. Como el único $a_{i,3} > 0$ es el correspondiente a la primera restricción se tiene que la variable que saldrá de la base es x_5; así

	0	0	0	0	-1	0	-1	0	-15
x_2	$\frac{1}{2}$	0	1	-1	1	0	-$\frac{1}{2}$	0	$\frac{9}{2}$
x_6	2	0	0	0	1	1	0	0	9
x_2	$\frac{1}{2}$	1	0	0	0	0	$\frac{1}{2}$	0	$\frac{5}{2}$
x_8	1	0	0	0	0	0	0	1	4

Dado que todos los coeficientes de la función objetivo son menores o iguales que cero, el punto actual será punto factible óptimo, así el punto $x_1 = 0, x_2 = \frac{5}{2}, x_3 = \frac{9}{2}$, $x_4 = 0$, $x_5 = 0$, $x_6 = 0$, $x_7 = 0$, $x_8 = 4$ es punto factible óptimo, siendo el valor del problema $Z = 15$. En función de las variables de decisión el óptimo, se alcanza en el punto $(x, y, z) = (0, \frac{5}{2}, \frac{9}{2})$.

Ejemplo 3.3.7
Sea el problema de programación lineal dado por

$$\begin{aligned}&\text{Máx } Z = y - x\\&\text{sometido a:}\\&y - 2x \leq 1\\&2y - x \geq -1\end{aligned}$$

Puesto que no se indica que $x \geq 0$, $y \geq 0$ se introducen las nuevas variables $x_1, x_2, x_3, x_4 \geq 0$ de modo que $x = x_1 - x_2$, $y = x_3 - x_4$.
Introduciendo las variables de holgura $x_5, x_6 \geq 0$ se tiene el problema:

Máx $Z = -x_1 + x_2 + x_3 - x_4$ sometido a las restricciones:

$$\begin{aligned}-2x_1 + 2x_2 + x_3 - x_4 + x_5 &= 1\\x_1 - x_2 - 2x_3 + 2x_4 + x_6 &= 1\\x_1, x_2, x_3, x_4, x_5, x_6 &\geq 0\end{aligned}$$

lo que en forma tabular se expresa como:

-1	1	1	-1	0	0	0
-2	2	1	-1	1	0	1
1	-1	-2	2	0	1	1

Las variables básicas son x_5, x_6.

El mayor coeficiente positivo de la función objetivo es 1; por comodidad haremos entrar en la base x_3 puesto que su coeficiente en la primera ecuación es la unidad. Así, se tiene:

1	-1	0	0	-1	0	-1
-2	2	1	-1	1	0	1
-3	3	0	0	2	1	3

Puesto que el único coeficiente positivo de la función objetivo es el correspondiente a x_1 y los coeficientes de x_1 en las ecuaciones de restricción son negativos, se tiene que el problema tiene solución no acotada $Z = +\infty$. ♦

Ejemplo 3.3.8
Resolver el problema de programación lineal:

$$\begin{aligned}
&\text{máx}\, Z = x + y \quad \text{sometida a las restricciones:}\\
&2x + 2y \leq 1\\
&3x + 3y \geq 1\\
&x \geq \frac{1}{5}\\
&x, y \geq 0
\end{aligned}$$

Para resolver el problema, puesto que $x, y \geq 0$, hacemos $x = x_1$, $y = x_2$ e introducimos las variables de holgura $x_3, x_4, x_5 \geq 0$ para reducir el problema a su formulación estándar. Maximizar la función $Z = x_1 + x_2$ sometida a las restricciones:

$$\begin{aligned}
&2x_1 + 2x_2 + x_3 && = 1\\
&3x_1 + 3x_2 + \cdots - x_4 && = 1\\
&x_1 + \cdots \ldots\ldots\ldots - x_5 && = \frac{1}{5}\\
&x_1, x_2, x_3, x_4, x_5 \geq 0
\end{aligned}$$

Puesto que las columnas correspondientes a las variables de holgura no forman una base admisible, es necesario introducir variables artificiales para completar la base. Así, introduciendo las variables $x_6, x_7 \geq 0$ se obtiene el problema auxiliar correspondiente a la fase I:

Fase I:

Maximizar la función $Z = -x_6 - x_7$ sometida a las restricciones:

$$\begin{aligned}
&2x_1 + 2x_2 + x_3 + \cdots && = 1\\
&3x_1 + 3x_2 + \cdots - x_4 + \cdots + x_6 && = 1\\
&x + \cdots \ldots - x_5 + x_7 && = \frac{1}{5}\\
&x_1, x_2, x_3, x_4, x_5, x_6, x_7 \geq 0
\end{aligned}$$

lo que en forma tabular se escribe como:

0	0	0	0	0	-1	-1	0
2	2	1	0	0	0	0	1
3	3	0	-1	0	1	0	1
1	0	0	0	-1	0	1	$\frac{1}{5}$

Escribiendo la función objetivo como función de variables no básicas se tiene:

4	3	0	-1	-1	0	0	$\frac{6}{5}$
2	2	1	0	0	0	0	1
3	3	0	-1	0	1	0	1
1	0	0	0	-1	0	1	$\frac{1}{5}$

el mayor valor positivo de la función objetivo es el correspondiente a la primera columna; así:

4	3	0	-1	-1	0	0	$\frac{6}{5}$
1	1	$\frac{1}{2}$	0	0	0	0	$\frac{1}{2}$
1	1	0	-$\frac{1}{3}$	0	$\frac{1}{3}$	0	$\frac{1}{3}$
1	0	0	0	-1	0	1	$\frac{1}{5}$

el menor valor es el correspondiente a la tercera fila; así, saldrá de la base x_7 y entrará x_1.

0	3	0	-1	3	0	-4	$\frac{2}{5}$
0	1	$\frac{1}{2}$	0	1	0	-1	$\frac{3}{10}$
0	1	0	-$\frac{1}{3}$	1	$\frac{1}{3}$	-1	$\frac{2}{15}$
1	0	0	0	-1	0	1	$\frac{1}{5}$

El mayor valor de la función objetivo es el correspondiente a la quinta columna; así, se tiene:

0	0	0	0	0	-1	-1	0
0	0	$\frac{1}{2}$	$\frac{1}{3}$	0	-$\frac{1}{3}$	0	$\frac{1}{6}$
0	1	0	-$\frac{1}{3}$	1	$\frac{1}{3}$	-1	$\frac{2}{15}$
1	1	0	-$\frac{1}{3}$	0	$\frac{1}{3}$	0	$\frac{1}{3}$

con lo que finaliza la fase uno $x_6 = x_7 = 0$, $Z_{aux} = 0$.

Fase II:

1	1	0	0	0	0
0	0	$\frac{1}{2}$	$\frac{1}{3}$	0	$\frac{1}{6}$
0	1	0	-$\frac{1}{3}$	1	$\frac{2}{15}$
1	1	0	-$\frac{1}{3}$	0	$\frac{1}{3}$

Escribiendo la función objetivo en función de variables no básicas se obtiene

0	0	0	$\frac{1}{3}$	0	-$\frac{1}{3}$
0	0	$\frac{1}{2}$	$\frac{1}{3}$	0	$\frac{1}{6}$
0	1	0	-$\frac{1}{3}$	1	$\frac{2}{15}$
1	1	0	-$\frac{1}{3}$	0	$\frac{1}{3}$

puesto que el único valor positivo de la función objetivo es el correspondiente a la cuarta columna, se tiene:

0	0	0	$\frac{1}{3}$	0	-$\frac{1}{3}$
0	0	$\frac{3}{2}$	1	0	$\frac{1}{2}$
0	1	0	-$\frac{1}{3}$	1	$\frac{2}{15}$
1	1	0	-$\frac{1}{3}$	0	$\frac{1}{3}$

y por tanto:

0	0	-$\frac{1}{2}$	0	0	-$\frac{1}{2}$
0	0	$\frac{3}{2}$	1	0	$\frac{1}{2}$
0	1	$\frac{1}{2}$	0	1	$\frac{3}{10}$
1	1	$\frac{1}{2}$	0	0	$\frac{1}{2}$

Por tanto, una solución del problema es $x_1 = \frac{1}{2}$, $x_2 = x_3 = 0$, $x_4 = \frac{1}{2}$, $x_5 = \frac{3}{2}$, $Z = \frac{1}{2}$. ♦

Ejemplo 3.3.9
Obtener el mínimo de la función $Z = x - y + 4$ sometida a las restricciones

$$\begin{aligned} x + y + z &\leq 1 \\ x + 2y &\geq 7 \\ x, y, z &\geq 0 \end{aligned}$$

Puesto que $x, y, z \geq 0$ basará $x = x_1$, $y = x_2$, $z = x_3$ con $x_1, x_2, x_3 \geq 0$. Introduciendo variables de holgura $x_4, x_5 \geq 0$ se tiene

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 &= 1 \\ x_1 + 2x_2 - x_5 &= 7 \\ x_1, x_2, x_3, x_4, x_5 &\geq 0 \end{aligned}$$

Puesto que la variable de holgura x_5 aparece con signo menos, para formar base inicial es necesario introducir la variable artificial $x_6 \geq 0$, con lo que resulta

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 &= 1 \\ x_1 + 2x_2 - x_5 + x_6 &= 7 \\ x_1, x_2, x_3, x_4, x_5, x_6 &\geq 0 \end{aligned}$$

Fase I:
Se procede a maximizar la función auxiliar $Z_{aux} = -x_6$.
Expresando el problema en forma tabular se tiene

0	0	0	0	0	-1	0
1	1	1	1	0	0	1
1	2	0	0	-1	1	7

de donde

1	2	0	0	-1	0	7
1	1	1	1	0	0	1
1	2	0	0	-1	1	7

El mayor coeficiente positivo de la función objetivo es el correspondiente a la segunda columna; así

1	2	0	0	-1	0	7
1	1	1	1	0	0	1
$\frac{1}{2}$	1	0	0	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{7}{2}$

el menor término independiente es el correspondiente a la primera ecuación; así, se tiene

-1	0	-2	-2	-1	0	5
1	1	1	1	0	0	1
$-\frac{1}{2}$	0	-1	-1	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{5}{2}$

dado que todos los coeficientes de la función objetivo son negativos, se tiene que el punto actual es punto óptimo, siendo el valor del problema $Z_{aux} = -5$ el cual se alcanza en el punto $x_1 = 0$, $x_2 = 1$, $x_3 = 0$, $x_4 = 0$, $x_5 = 0$, $x_6 = 5$. Dado que dicho valor no es cero, se tiene que el conjunto factible del problema inicial es vacío y, por tanto, no hay puntos óptimos. ♦

Solución general

En ocasiones es posible llevar el problema de programación lineal a una forma en la que las variables no básicas vienen dadas por $x_{i_1}, \ldots, x_r$ y en la que la función objetivo se expresa como $c_{i_1}x_{i_1}+\cdots+c_{i_s}x_{i_s}+d$ con $s < r$ y donde $c_{i_1} < 0$,,. . .,$c_{i_s} < 0$. En este caso, para obtener el mínimo de la función objetivo podemos proceder a igualar a cero las variables no básicas, y resolver el sistema restante para las básicas, con lo que obtiene un punto factible óptimo en el cual se alcanza el valor del problema $Z = d$.

En este caso, puesto que la función objetivo no depende de $x_{i_{s+1}}, \ldots, x_{i_k}$, para obtener dicho máximo bastará con anular las variables no básicas que aparecen en la función objetivo, alcanzándose también el valor $Z = d$, pues esta es independiente del valor que tomen las variables $x_{i_{s+1}}, \ldots, x_{i_k}$. Así, el valor del problema se obtendrá en el conjunto

$$S = \{\overrightarrow{x} \in \mathbb{R}^n : A\overrightarrow{x} = \overrightarrow{b}, \overrightarrow{x} > \overrightarrow{0}, x_{i_1} = 0, \ldots, x_{i_s} = 0\}.$$

Este conjunto es conveniente escribirlo partiendo de la formulación estándar inicial, pues siempre es deseable dar la respuesta del problema en función de las variables de decisión.

Ejemplo 3.3.10

$$\begin{aligned}&\text{Maximizar } Z = x + y\\&\text{sometido a:}\\&3x + 3y \geq 1\\&2x + 2y \leq 1\\&x, y \geq 0\end{aligned} \tag{3.3}$$

Encontrar su solución general.

Para expresar el problema en función de variables positivas, dado que $x, y \geq 0$ basta definir $x = x_1$, e $y = x_2$. A continuación se introducen las variables de holgura $x_3, x_4 \geq 0$ de modo que el problema se reescribe como:
Máx $Z = x_1 + x_2$ sometido a las restricciones:

$$\begin{aligned}3x_1 + 3x_2 - x_3 &= 1\\2x_1 + 2x_2 + x_4 &= 1\\x_1, x_2, x_3, x_4 &\geq 0\end{aligned}$$

Fase I:

Puesto que la variable de holgura x_3 aparece con signo menos será necesario introducir la variable artificial x_5; así, se tiene

$$\begin{aligned} 3x_1 + 3x_2 - x_3 + x_5 &= 1 \\ 2x_1 + 2x_2 + x_4 &= 1 \\ x_1, x_2, x_3, x_4 &\geq 0 \end{aligned}$$

lo que en forma tabular se escribe como:

0	0	0	0	-1	0
3	3	-1	0	1	1
2	2	0	1	0	1

3	3	-1	0	0	1
3	3	-1	0	1	1
2	2	0	1	0	1

3	3	-1	0	0	1
1	1	$-\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{1}{3}$
1	1	0	$\frac{1}{2}$	0	$\frac{1}{2}$

Tomando como nueva variable básica x_1, se tiene:

0	0	0	0	-1	0
1	1	$-\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{1}{3}$
0	0	$\frac{1}{3}$	$\frac{1}{2}$	$-\frac{1}{3}$	$\frac{1}{6}$

con lo que finaliza la fase I.

Fase II:

1	1	0	0	0
1	1	$-\frac{1}{3}$	0	$\frac{1}{3}$
0	0	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{6}$

Para tener las columnas de la base x_1, x_4 en forma canónica, se multiplica por dos la última fila, así

1	1	0	0	0
1	1	$-\frac{1}{3}$	0	$\frac{1}{3}$
0	0	$\frac{2}{3}$	1	$\frac{1}{3}$

Expresando la función objetivo en función de variables no básicas se tiene

0	0	$\frac{1}{3}$	0	$-\frac{1}{3}$
1	1	$-\frac{1}{3}$	0	$\frac{1}{3}$
0	0	$\frac{2}{3}$	1	$\frac{1}{3}$

por tanto:

$$\begin{array}{cccc|c} 0 & 0 & \frac{1}{3} & 0 & -\frac{1}{3} \\ \hline 1 & 1 & -\frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & \frac{3}{2} & \frac{1}{2} \end{array}$$

de donde

$$\begin{array}{cccc|c} 0 & 0 & 0 & -\frac{1}{2} & -\frac{1}{2} \\ \hline 1 & 1 & 0 & -\frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 & \frac{3}{2} & \frac{1}{2} \end{array}$$

Las variables básicas son x_1, x_3; así la solución general del problema lineal asociado es que la función Z alcanza un valor máximo de $\frac{1}{2}$. Puesto que la variable no básica x_2 no aparece en la función objetivo no es necesario anularla; así, partiendo de la forma estándar inicial, se tiene que el conjunto S donde se alcanza el máximo está dado por las ecuaciones

$$\begin{aligned} 3x_1 + 3x_2 - x_3 &= 1 \\ 2x_1 + 2x_2 + x_4 &= 1 \\ x_1, x_2, x_3, x_4 &\geq 0 \end{aligned}$$

a la que se añade $x_4 = 0$; así, resulta

$$\begin{aligned} 3x_1 + 3x_2 - x_3 &= 1 \\ 2x_1 + 2x_2 &= 1 \\ x_1, x_2, x_3 &\geq 0 \end{aligned}$$

eliminando las variables de holgura podemos escribir en forma de desigualdades

$$\begin{aligned} 3x_1 + 3x_2 &\leq 1 \\ 2x_1 + 2x_2 &= 1 \\ x_1, x_2 &\geq 0 \end{aligned}$$

cuya solución es:

$$\begin{aligned} 2x_1 + 2x_2 &= 1 \\ x_1, x_2 &\geq 0 \end{aligned}$$

Escribiendo este resultado en función de las variables de decisión se tiene

$$\begin{aligned} 2x + 2y &= 1 \\ x, y &\geq 0 \end{aligned}$$

$$S = \left\{ (x, y) \in \mathbb{R}^2 | x + y = \frac{1}{2}, x \geq 0, y \geq 0 \right\}$$

♦

Tema 4

Programación entera

4.1. Introducción

El problema de programación entera es un problema consistente en optimizar $f(x_1, \ldots, x_n)$, sujeto a restricciones dadas por $g_i(x_1, \ldots, x_n) \leq b_i$, $i = 1, \ldots, m$ con $x_i \geq 0$, $i = 1, \ldots, n$, y $x_j \in \mathbb{Z}^+ = \{0, 1, 2, \ldots\}$, $\forall j \in I \subset \{1, \ldots, n\}$. Cuando $I = \{1, \ldots, n\}$ diremos que el problema es entero puro y cuando $I \neq \{1, \ldots, n\}$ diremos que el problema es entero mixto. El problema de programación entera más estudiado es el lineal, el cual podemos formular como:

Optimizar $< \overrightarrow{c}, \overrightarrow{x} >$, sometido a $A\overrightarrow{x} \leq \overrightarrow{b}$, $A \in \mathbb{R}_{m \times n}$, $\overrightarrow{b} \in \mathbb{R}^m$, $\overrightarrow{b} \geq \overrightarrow{0}$.

$x_j \geq 0$, $j \in \{1, \ldots, n\}$; $x_j \in \mathbb{Z}^+, \forall j \in I \subset \{1, \ldots, n\}$.

Nuestro estudio se reducirá, en este curso, a problemas de programación lineal entera (en particular cuando $I = \{1, \ldots, n\}$).
Cuando se intenta la resolución de un problema de programación entera, parece lógico pensar que un buen procedimiento puede ser la resolución del problema sin contar con la restricción de variables enteras, para a continuación a partir de la solución del problema continuo buscar por aproximación por redondeo la solución del caso discreto. Lamentablemente, este procedimiento no proporciona en muchos casos buenos resultados.
Los procedimientos principales para la resolución del problema de programación entera se clasifican en métodos algebraicos, métodos combinatorios y métodos de enumeración. Estos métodos consisten fundamentalmente en:

Métodos algebraicos: consisten en añadir restricciones al problema continuo asociado de modo que éste, junto a las nuevas restricciones, tenga solución factible óptima entera. Este método consiste en eliminar de la región factible del problema continuo partes que no son factibles en el problema discreto. También recibe el nombre de método de los conjuntos de corte.

Métodos combinatorios: en este grupo se incluyen algoritmos de naturaleza combinatoria que poseen la propiedad de que el número de

pasos está acotado por una expresión de tipo polinómica en n. Entre estos métodos está el de Goromy, consistente en ir efectuando redondeos a las variables del problema continuo, imponiendo a su vez condiciones adicionales, lo que conduce a un proceso de programación dinámica.

Métodos de enumeración: partiendo de que el número de posibles soluciones óptimas debe ser finito, se busca mediante la enumeración de éstos (o exploración dirigida) una solución factible óptima. Este método suele utilizarse en un tipo especial de problemas de programación entera; los problemas cero-uno, en los que las variables pueden tomar únicamente valor cero o uno.

En el presente curso únicamente estudiaremos en cierta profundidad los métodos algebraicos, en particular los métodos de ramificación, y el método de los planos de cortes, lo cual es suficiente para nuestros propósitos. Los métodos de ramificación y de planos de corte parten de la solución del problema lineal asociado, problema que a continuación definimos:

Definición 4.1.1
Sea el problema de programación entera definido por:

Optimizar $< \overrightarrow{c}, \overrightarrow{x} >$,

sometido a $A\overrightarrow{x} \leq \overrightarrow{b}, A \in \mathbb{R}_{m\times n}, \overrightarrow{b} \in \mathbb{R}^m, \overrightarrow{b} \geq \overrightarrow{0}$.

$x_j \geq 0, j \in \{1, \ldots, n\}; x_j \in Z^+, \forall j \in I \subset \{1, \ldots, n\}$.

Llamamos **programa lineal asociado** al problema:

Optimizar $Z =< \overrightarrow{c}, \overrightarrow{x} > +d$.

sometido a $A\overrightarrow{x} \leq \overrightarrow{b}, A \in \mathbb{R}_{m\times n}, d \in \mathbb{R}$.

$x_j \geq 0, j \in \{1, \ldots, n\}$.

4.2. Método de ramificación

Sea el problema de programación entera dado por

Optimizar $< \overrightarrow{c}, \overrightarrow{x} >$,

sometido a $A\overrightarrow{x} \leq \overrightarrow{b}, A \in \mathbb{R}_{m\times n}, \overrightarrow{b} \in \mathbb{R}^m, \overrightarrow{b} \geq 0$.

$x_j \geq 0, j \in \{1, \ldots, n\}; x_j \in \mathbb{Z}^+, \forall j \in \{1, \ldots, r\}, r \leq n$.

El método de ramificación consiste en resolver, en primer lugar, el problema lineal asociado, obteniéndose para éste una solución $\overrightarrow{x}$. Entonces si $j \in \{1, \ldots, r\}$ y $x_j = t_j \notin \mathbb{Z}$ resulta que en el problema entero se debe verificar necesariamente una de las siguientes condiciones: $x_j \leq E[t_j]$ o bien $x_j \geq E[t_j] + 1$, donde $E[\ ]$ representa la función parte entera. Consecuencia de esto es que el problema inicial se descompone en una serie de problemas consistentes en añadir a éste las restricciones derivadas de las condiciones $x_j \leq E[t_j]$, $x_j \geq E[t_j] + 1$ donde

$j \in \{i \in \{1, \ldots, r\} : t_j \notin \mathbb{Z}\}$. Nótese que esto implica resolver 2^p problemas donde p es el número de variables no enteras.

La solución del problema entero se alcanza cuando, una vez resueltos estos problemas, resulta que el mayor de los máximos de los distintos problemas se alcanza en un punto en el que se cumplen las condiciones de integridad, en cuyo caso, ésta es la solución del problema. En caso de no cumplirse lo anterior, debe procederse a la ramificación de los problemas anteriores comenzando por el de mayor máximo hasta que se satisfaga la condición de integridad requerida.

Ejemplo 4.2.1

Maximizar la función

$Z = x_1 + 2x_2$

sometida a las restricciones

$4x_1 + 3x_2 \leq 12$

$-x_1 + \quad x_2 \leq \ \ 2$

$x_1, x_2 \geq 0$

$x_1, x_2 \in \mathbb{Z}$

En primer lugar, se tiene que el problema lineal asociado viene dado por

Maximizar la función

$Z = x_1 + 2x_2$

sometida a las restricciones

$4x_1 + 3x_2 \leq 12$

$-x_1 + \quad x_2 \leq \ \ 2$

$x_1, x_2 \geq 0$

que en su forma estándar resulta

Maximizar $Z = x_1 + 2x_2$

sometida a las restricciones

$$\begin{array}{rl} 4x_1 + 3x_2 + x_3 \qquad & = 12 \\ - \ \ x_1 + \ \ x_2 \qquad + x_4 & = \ \ 2 \end{array}$$

$x_1, x_2, x_3, x_4 \geq 0$

lo que en forma tabular resulta

Básicas	1	2	0	0	0
x_3	4	3	1	0	12
x_4	-1	1	0	1	2
	0	0	12	2	

El mayor coeficiente de la función objetivo es el correspondiente a la variable x_2; así se tiene

Básicas	1	2	0	0	0
x_3	$\frac{4}{3}$	1	$\frac{1}{3}$	0	4
x_4	-1	1	0	1	2

y puesto que el menor término independiente correspondiente a las filas en que es uno, el coeficiente de x_2 es el de la segunda, se tiene

Básicas	3	0	0	-2	-4
x_3	$\frac{7}{3}$	0	$\frac{1}{3}$	-1	2
x_2	-1	1	0	1	2

El mayor coeficiente positivo de la función objetivo es el correspondiente a la variable x_1; así, se tiene

Básicas	3	0	0	-2	-4
x_3	1	0	$\frac{1}{7}$	$-\frac{3}{7}$	$\frac{6}{7}$
x_2	-1	1	0	1	2

de donde

Básicas	0	0	$-\frac{3}{7}$	$-\frac{5}{7}$	$-\frac{46}{7}$
x_1	1	0	$\frac{1}{7}$	$-\frac{3}{7}$	$\frac{6}{7}$
x_2	0	1	$\frac{1}{7}$	$\frac{4}{7}$	$\frac{20}{7}$
	$\frac{46}{7}$	$\frac{20}{7}$	0	0	

Puesto que la función objetivo está expresada en función de variables no básicas y todos sus coeficientes son negativos, se tiene que la solución actual es máximo de Z. Dicha solución viene dada por $x_1 = \frac{6}{7}$, $x_2 = \frac{20}{7}$, $x_3 = 0$, $x_4 = 0$. Esta solución no cumple el requisito de integridad establecido, por lo cual deben imponerse las condiciones adicionales para x_1 de cumplir $x_1 \leq 0$ ó $x_1 \geq 1$, y para x_2, $x_2 \leq 2$ ó $x_3 \geq 3$.

De este modo tenemos cuatro posibles condiciones adicionales dadas por:

- Caso I: $x_1 \leq 0$, $x_2 \leq 2$.
- Caso II: $x_1 \leq 0$, $x_3 \geq 3$.
- Caso III: $x_1 \geq 1$, $x_2 \leq 2$.
- Caso IV: $x_1 \geq 1$, $x_2 \geq 3$.

A continuación procedemos a resolver estos casos de modo independiente.

Caso I: En este caso necesariamente $x_1 = 0$, por lo que el problema se podría simplificar. No obstante, será resuelto sin tener en cuenta dicha circunstancia.

$$\begin{aligned}
&\text{máx}\, Z = x_1 + 2x_2 \\
&\text{sometida a las restricciones} \\
&4x_1 + 3x_2 \leq 12 \\
&-x_1 + \quad x_2 \leq \;\; 2 \\
&x_1 \leq 0 \\
&x_2 \leq 2 \\
&x_1, x_2 \geq 0
\end{aligned}$$

que en su forma estándar resulta

$$\begin{aligned}
&\text{Maximizar } Z = x_1 + 2x_2 \\
&\text{sometida a las restricciones} \\
&\;4x_1 + 3x_2 + x_3 \qquad\qquad\qquad = 12 \\
&-x_1 + \;\; x_2 \quad\;\; + x_4 \qquad\quad\; = \;\; 2 \\
&\;\; x_1 + \qquad\qquad\qquad + x_5 \quad\; = \;\; 0 \\
&\qquad + \;\; x_2 \qquad\qquad\qquad + x_6 = \;\; 2 \\
&x_1, x_2, x_3, x_4, x_5, x_6 \geq 0
\end{aligned}$$

lo que en forma tabular resulta

1	2	0	0	0	0	0
4	3	1	0	0	0	12
-1	1	0	1	0	0	2
1	0	0	0	1	0	0
0	1	0	0	0	1	2

así

1	2	0	0	0	0	0
$\frac{4}{3}$	1	$\frac{4}{3}$	0	0	0	4
-1	1	0	1	0	0	2
1	0	0	0	1	0	0
0	1	0	0	0	1	2

de donde

3	0	0	-2	0	0	-4
$\frac{7}{3}$	0	$\frac{1}{3}$	-1	0	0	2
-1	1	0	1	0	0	2
1	0	0	0	1	0	0
1	0	0	-1	0	1	0

expresando la base en forma canónica se tiene

3	0	0	-2	0	0	-4
7	0	1	-3	0	0	6
-1	1	0	1	0	0	2
1	0	0	0	1	0	0
1	0	0	-1	0	1	0

a continuación, se tiene

0	0	0	-2	-1	0	-4
7	0	1	-3	-7	0	6
0	1	0	1	1	0	2
1	0	0	0	1	0	0
0	0	0	-1	-1	1	0

Dado que todos los coeficientes de la función objetivo son menores que cero, se tiene que el punto actual $x_1 = 0$, $x_2 = 2$, $x_3 = 6$, $x_4 = 0$, $x_5 = 0$, $x_6 = 0$ es punto factible óptimo y dado que cumple la condición de integridad, es solución del problema entero en el caso I.

El valor del problema resulta $Z_I = 4$.

Caso II:

$$
\begin{aligned}
&\text{máx}\, Z = x_1 + 2x_2 \\
&\text{sometida a las restricciones} \\
&4x_1 + 3x_2 \leq 12 \\
&-x_1 + \quad x_2 \leq \quad 2 \\
&x_1 \leq 0 \\
&x_2 \geq 3 \\
&x_1, x_2 \geq 0
\end{aligned}
$$

que en su forma estándar resulta

$$
\begin{aligned}
&\text{Maximizar } Z = x_1 + 2x_2 \\
&\text{sometida a las restricciones} \\
&\quad 4x_1 + 3x_2 + x_3 \qquad\qquad = 12 \\
&- \quad x_1 + \quad x_2 \qquad\qquad = \quad 2 \\
&\quad x_1 + \qquad\qquad + x_5 \quad = \quad 0 \\
&\qquad + \quad x_2 \qquad\qquad - x_6 = \quad 3 \\
&x_1, x_2, x_3, x_4, x_5, x_6 \geq 0
\end{aligned}
$$

Puesto que la variable de holgura x_6 aparece con signo menos, será necesario buscar, en primer lugar, una base inicial, lo que trataremos de lograr mediante la introducción de una variable artificial y maximizando la función auxiliar $Z_{aux} = -x_7$. El problema resultante lo podemos escribir en forma tabular como:

Fase I

0	0	0	0	0	0	-1	0
4	3	1	0	0	0	0	12
-1	1	0	1	0	0	0	2
1	0	0	0	1	0	0	0
0	1	0	0	0	-1	1	3

Expresando la función objetivo en función de variables no básicas, se tiene

0	1	0	0	0	-1	0	3
4	3	1	0	0	0	0	12
-1	1	0	1	0	0	0	2
1	0	0	0	1	0	0	0
0	1	0	0	0	-1	1	3

Aplicando el método del simplex se tiene

0	1	0	0	0	-1	0	3
$\frac{4}{3}$	1	$\frac{1}{3}$	0	0	0	0	4
-1	1	0	1	0	0	0	2
1	0	0	0	1	0	0	0
0	1	0	0	0	-1	1	3

de donde

0	1	0	0	0	-1	0	3
$\frac{4}{3}$	1	$\frac{1}{3}$	0	0	0	0	4
-1	1	0	1	0	0	0	2
1	0	0	0	1	0	0	0
0	1	0	0	0	-1	1	3

A continuación

1	0	0	-1	0	0	0	1
$\frac{7}{3}$	0	$\frac{1}{3}$	1	0	0	0	2
-1	1	0	1	0	0	0	2
1	0	0	0	1	0	0	0
1	0	0	-1	0	-1	1	1

expresando la base en forma canónica, se tiene

1	0	0	-1	0	0	0	1
7	0	1	3	0	0	0	6
-1	1	0	1	0	0	0	2
1	0	0	0	1	0	0	0
1	0	0	-1	0	-1	1	1

A continuación se tiene

1	0	0	-1	0	0	0	1
1	0	$\frac{1}{7}$	$\frac{3}{7}$	0	0	0	$\frac{6}{7}$
-1	1	0	1	0	0	0	2
1	0	0	0	1	0	0	0
1	0	0	-1	0	-1	1	1

de donde

0	0	0	-1	-1	0	0	1
1	0	$\frac{1}{7}$	$\frac{3}{7}$	0	0	0	$\frac{6}{7}$
-1	1	0	1	0	0	0	2
1	0	0	0	1	0	0	0
1	0	0	-1	0	-1	1	1

Finalmente

0	0	0	-1	-1	0	0	1
1	0	$\frac{1}{7}$	$\frac{3}{7}$	0	0	0	$\frac{6}{7}$
-1	1	0	1	0	0	0	2
1	0	0	0	1	0	0	0
1	0	0	-1	0	-1	1	1

En este caso, se obtiene $Z_{aux} = -1$ con lo que el conjunto factible del problema correspondiente al caso II es vacío.

Caso III

$$\begin{aligned}
&\text{máx}\, Z = x_1 + 2x_2\\
&\text{sometida a las restricciones}\\
&4x_1 + 3x_2 \leq 12\\
&-x_1 + \quad x_2 \leq \ \ 2\\
&x_1 \geq 1\\
&x_2 \leq 2\\
&x_1, x_2 \geq 0
\end{aligned}$$

que en su forma estándar resulta

$$\begin{aligned}
&\text{Maximizar } Z = x_1 + 2x_2\\
&\text{sometida a las restricciones}\\
&\ \ 4x_1 + 3x_2 + x_3 \qquad\qquad\qquad = 12\\
&-\ \ x_1 + \ \ x_2 \qquad\qquad\qquad\quad = \ \ 2\\
&\ \ \ x_1 + \qquad\qquad\qquad - x_5 \quad = \ \ 1\\
&\qquad + \ \ x_2 \qquad\qquad\qquad + x_6 = \ \ 2\\
&x_1, x_2, x_3, x_4, x_5, x_6 \geq 0
\end{aligned}$$

Puesto que la variable de holgura x_5 aparece con signo menos, será necesario buscar, en primer lugar, una base inicial, lo que trataremos de lograr mediante la introducción de una variable artificial $x_7 \geq 0$ y maximizando la función auxiliar $Z_{aux} = -x_7$. El problema resultante lo podemos escribir en forma tabular como:

Fase I:

0	0	0	0	0	0	-1	0
4	3	1	0	0	0	0	12
-1	1	0	1	0	0	0	2
1	0	0	0	-1	0	1	1
0	1	0	0	0	1	0	2

Expresando la función objetivo en función de variables no básicas, se tiene

1	0	0	0	-1	0	0	1
4	3	1	0	0	0	0	12
-1	1	0	1	0	0	0	2
1	0	0	0	-1	0	1	1
0	1	0	0	0	1	0	2

Aplicando el método del simplex, se tiene

1	0	0	0	-1	0	0	1
1	$\frac{3}{4}$	$\frac{1}{4}$	0	0	0	0	3
-1	1	0	1	0	0	0	2
1	0	0	0	-1	0	1	1
0	1	0	0	0	1	0	2

y de aquí

0	0	0	0	0	0	-1	0
0	$\frac{3}{4}$	$\frac{1}{4}$	0	1	0	-1	2
0	1	0	1	-1	0	1	3
1	0	0	0	-1	0	1	1
0	1	0	0	0	1	0	2

Puesto que la función objetivo auxiliar alcanza su máximo $Z_{aux} = 0$ en $x_7 = 0$ se pasa a la fase II.

Fase II

1	2	0	0	0	0	0
0	$\frac{3}{4}$	$\frac{1}{4}$	0	1	0	2
0	1	0	1	-1	0	3
1	0	0	0	-1	0	1
0	1	0	0	0	1	2

Escribiendo la función objetivo en función de variables no básicas, se tiene

0	2	0	0	1	0	-1
0	$\frac{3}{4}$	$\frac{1}{4}$	0	1	0	2
0	1	0	1	-1	0	3
1	0	0	0	-1	0	1
0	1	0	0	0	1	2

de donde

0	2	0	0	1	0	-1
0	1	$\frac{1}{3}$	0	$\frac{4}{3}$	0	$\frac{8}{3}$
0	1	0	1	-1	0	3
1	0	0	0	-1	0	1
0	1	0	0	0	1	2

de donde

0	0	0	0	1	-2	-5
0	0	$\frac{1}{3}$	0	$\frac{4}{3}$	-1	$\frac{2}{3}$
0	0	0	1	-1	-1	3
1	0	0	0	-1	0	1
0	1	0	0	0	1	2

por tanto,

0	0	0	0	1	-2	-5
0	0	$\frac{1}{4}$	0	1	-$\frac{3}{4}$	$\frac{1}{2}$
0	0	0	1	-1	-1	3
1	0	0	0	-1	0	1
0	1	0	0	0	1	2

Así, resulta

0	0	-$\frac{1}{4}$	0		-$\frac{5}{4}$	-$\frac{11}{2}$
0	0	$\frac{1}{4}$	0	1	-$\frac{3}{4}$	$\frac{1}{2}$
0	0	$\frac{1}{4}$	1	0	-$\frac{7}{4}$	$\frac{3}{2}$
1	0	$\frac{1}{4}$	0	0	-$\frac{3}{4}$	$\frac{3}{2}$
0	1	0	0	0	1	2

El punto óptimo se encuentra en $x_1 = \frac{3}{2}$, $x_2 = 2$, $x_4 = x_6 = 0$, $x_3 = \frac{3}{2}$, $x_4 = \frac{1}{2}$, siendo el valor del problema $Z = \frac{11}{2}$.

Caso IV

$$
\begin{aligned}
&\text{máx}\, Z = x_1 + 2x_2 \\
&\text{sometida a las restricciones} \\
&4x_1 + 3x_2 \leq 12 \\
&-x_1 + \quad x_2 \leq \ \ 2 \\
&x_1 \geq 1 \\
&x_2 \geq 3 \\
&x_1, x_2 \geq 0
\end{aligned}
\tag{4.1}
$$

Para resolver este problema lo haremos directamente en forma tabular, siendo en este caso necesario recurrir a las dos fases; así, se tiene:

Fase I

0	0	0	0	0	0	-1	-1	0
4	3	1	0	0	0	0	0	12
-1	1	0	1	0	0	0	0	2
1	0	0	0	-1	0	1	0	1
0	1	0	0	0	-1	0	1	3

Expresando la función objetivo en función de variables no básicas, se tiene

1	1	0	0	-1	-1	0	0	4
4	3	1	0	0	0	0	0	12
-1	1	0	1	0	0	0	0	2
1	0	0	0	-1	0	1	0	1
0	1	0	0	0	-1	0	1	3

A continuación

1	1	0	0	-1	-1	0	0	4
1	$\frac{3}{4}$	$\frac{1}{4}$	0	0	0	0	0	3
-1	1	0	1	0	0	0	0	2
1	0	0	0	-1	0	1	0	1
0	1	0	0	0	-1	0	1	3

por tanto

1	1	0	0	-1	-1	0	0	4
1	$\frac{3}{4}$	$\frac{1}{4}$	0	0	0	0	0	3
-1	1	0	1	0	0	0	0	2
1	0	0	0	-1	0	1	0	1
0	1	0	0	0	-1	0	1	3

y así

0	1	0	0	0	-1	-1	0	3
0	$\frac{3}{4}$	$\frac{1}{4}$	0	1	0	-1	0	2
0	1	0	1	-1	0	1	0	3
1	0	0	0	-1	0	1	0	1
0	1	0	0	0	-1	0	1	3

A partir de esta tabla, se tiene

0	1	0	0	0	-1	-1	0	3
0	1	$\frac{1}{3}$	0	$\frac{4}{3}$	0	$-\frac{4}{3}$	0	$\frac{8}{3}$
0	1	0	1	-1	0	1	0	3
1	0	0	0	-1	0	1	0	1
0	1	0	0	0	-1	0	1	3

de donde

0	0	$-\frac{1}{3}$	0	$-\frac{4}{3}$	-1	$\frac{1}{3}$	0	$\frac{1}{3}$
0	1	$\frac{1}{3}$	0	$\frac{4}{3}$	0	$-\frac{4}{3}$	0	$\frac{8}{3}$
0	0	$-\frac{1}{3}$	1	$-\frac{7}{3}$	0	$\frac{7}{3}$	0	$\frac{1}{3}$
1	0	0	0	-1	0	1	0	1
0	0	$-\frac{1}{3}$	0	$-\frac{4}{3}$	-1	$\frac{4}{3}$	1	$\frac{1}{3}$

El mayor coeficiente positivo de Z_{aux} es $\frac{1}{3}$; así

0	0	$-\frac{1}{3}$	0	$-\frac{4}{3}$	-1	$\frac{1}{3}$	0	$\frac{1}{3}$
0	1	$\frac{1}{3}$	0	$\frac{4}{3}$	0	$-\frac{4}{3}$	0	$\frac{8}{3}$
0	0	$-\frac{1}{7}$	$\frac{3}{7}$	-1	0	1	0	$\frac{1}{7}$
1	0	0	0	-1	0	1	0	1
0	0	$-\frac{1}{4}$	0	-1	$-\frac{3}{4}$	1	$\frac{3}{4}$	$\frac{1}{4}$

Por tanto, en este caso el conjunto factible resulta vacío por no ser cero el máximo de la función objetivo auxiliar.

Análisis de resultados

En los casos anteriores se han obtenido los resultados

Caso:	I	II	III	VI
Z_{max}	4	$\emptyset$	$\frac{11}{2}$	$\emptyset$
x_{max}	0	$\emptyset$	$\frac{3}{2}$	$\emptyset$
y_{max}	2	$\emptyset$	2	$\emptyset$

El máximo se alcanza en el caso III, no siendo alcanzado con $x_1, x_2 \in \mathbb{Z}$, pues la variable x_1 no cumple la condición de integridad; por tanto, deberá separarse en dos regiones, $x_1 \leq 1$ y $x_1 \geq 2$, lo que se analizará como caso V cuando $x_1 \geq 1$, $x_1 \leq 1$, y $x_2 \leq 2$ y caso VI cuando $x_1 \geq 1$, $x_1 \geq 2$, $x_2 \leq 2$, lo que resulta equivalente a:

- Caso V: $x_1 = 1, x_2 \leq 2$.
- Caso VI: $x_1 \geq 2$ y $x_2 \leq 2$.

Caso V

En este caso procederemos a simplificar el problema (podría procederse igual sin hacer esto); así se tiene el problema inicial con las restricciones adicionales $x_1 = 1$, $x_2 \leq 2$, por tanto, este se reduce a :

$$\begin{aligned}
&\text{máx}\, Z = 1 + 2x_2 \\
&\text{sometida a las restricciones} \\
&4 + 3x_2 \leq 12 \\
&-1 + \quad x_2 \leq \;\; 3 \\
&x_2 \leq 2 \\
&x_2 \geq 0
\end{aligned} \tag{4.2}$$

lo que equivale a

$$\begin{aligned}
&\text{máx}\, Z = 2x_2 + 1 \\
&\text{sometida a las restricciones} \\
&x_2 \leq 2 \\
&x_2 \geq 0
\end{aligned}$$

Introduciendo variables de holgura y escribiendo en forma tabular, se tiene

$$\begin{array}{c|cc|c} & 2 & 0 & -1 \\ \hline x_3 & 1 & 1 & 2 \end{array} \rightarrow \begin{array}{c|ccc|} & 0 & -2 & -5 \\ \hline x_2 & 1 & 1 & 2 \end{array}$$

por tanto, $x_2 = 2$, $x_3 = 0$, $Z = 5$, y recordemos que $x_1 = 1$.

Caso VI

En este caso deben añadirse al problema inicial las condiciones $x_1 \geq 2$ y $x_2 \leq 2$.

$$\begin{array}{l} \text{máx } Z = x_1 + 2x_2 \\ \text{sometida a las restricciones} \\ 4x_1 + 3x_2 \leq 12 \\ -x_1 + \quad x_2 \leq \ 2 \\ x_1 \geq 2 \\ x_2 \leq 2 \\ x_1, x_2 \geq 0 \end{array}$$

que en su forma estándar resulta

$$\begin{array}{l} \text{Maximizar } Z = x_1 + 2x_2 \\ \text{sometida a las restricciones} \\ \quad 4x_1 + 3x_2 + x_3 \qquad\qquad\qquad = 12 \\ -\ \ x_1 + \ \ x_2 \qquad\qquad\qquad = \ \ 2 \\ \quad x_1 + \qquad\qquad\qquad - x_5 \qquad = \ \ 2 \\ \qquad + \ \ x_2 \qquad\qquad\qquad + x_6 = \ \ 2 \\ x_1, x_2, x_3, x_4, x_5, x_6 \geq 0 \end{array}$$

Puesto que la variable de holgura x_5 aparece con signo menos, será necesario buscar, en primer lugar, una base inicial, lo que trataremos de lograr mediante la introducción de una variable artificial $x_7 \geq 0$ y maximizando la función auxiliar $Z_{aux} = -x_7$. El problema resultante lo podemos escribir en forma tabular como:

Fase I:

0	0	0	0	0	0	-1	0
4	3	1	0	0	0	0	12
-1	1	0	1	0	0	0	2
1	0	0	0	-1	0	1	2
0	1	0	0	0	1	0	2

Expresando la función objetivo en función de variables no básicas, se obtiene

1	0	0	0	-1	0	0	2
4	3	1	0	0	0	0	12
-1	1	0	1	0	0	0	2
1	0	0	0	-1	0	1	2
0	1	0	0	0	1	0	2

por tanto,

1	0	0	0	-1	0	0	2
1	$\frac{3}{4}$	$\frac{1}{4}$	0	0	0	0	3
-1	1	0	1	0	0	0	2
1	0	0	0	-1	0	1	2
0	1	0	0	0	1	0	2

de donde

0	0	0	0	0	-1	0	
0	$\frac{3}{4}$	$\frac{1}{4}$	0	1	0	-1	1
0	1	0	1	-1	0	1	4
1	0	0	0	-1	0	1	2
0	1	0	0	0	1	0	2

por tanto Z_{aux} alcanza el valor $Z_{aux} = 0$ con $x_7 = 0$.

Fase II

1	2	0	0	0	0	0
0	$\frac{3}{4}$	$\frac{1}{4}$	0	1	0	1
0	1	0	1	-1	0	4
1	0	0	0	-1	0	2
0	1	0	0	0	1	2

Expresando la función objetivo en función de variables no básicas, se tiene

0	2	0	0	1	0	-2
0	$\frac{3}{4}$	$\frac{1}{4}$	0	1	0	1
0	1	0	1	-1	0	4
1	0	0	0	-1	0	2
0	1	0	0	0	1	2

El mayor coeficiente positivo de Z es el correspondiente a x_2, por tanto,

0	2	0	0	1	0	-2
0	1	$\frac{1}{3}$	0	$\frac{4}{3}$	0	$\frac{4}{3}$
0	1	0	1	-1	0	4
1	0	0	0	-1	0	2
0	1	0	0	0	1	2

de donde resulta

0	0	-$\frac{2}{3}$	0	-$\frac{5}{3}$	0	-$\frac{14}{3}$
0	1	$\frac{1}{3}$	0	$\frac{4}{3}$	0	$\frac{4}{3}$
0	0	-$\frac{1}{3}$	1	$-\frac{7}{3}$	0	$\frac{8}{3}$
1	0	0	0	-1	0	2
0	0	-$\frac{1}{3}$	0	-$\frac{4}{3}$	1	$\frac{2}{3}$

el máximo se alcanza para $x_1 = 2$, $x_2 = \frac{4}{3}$, $x_3 = x_5 = 0$, $x_4 = \frac{8}{3}$, $x_6 = \frac{2}{3}$, siendo el valor del problema $Z = \frac{14}{3}$.

Análisis de resultados

Una vez resueltos los casos V y VI, los cuales sustituyen al caso III. resulta la tabla:

Caso	I	II	VI	V	VI
Z_{max}	4	$\emptyset$	$\emptyset$	5	$\frac{14}{3}$
x_{max}	0	$\emptyset$	$\emptyset$	1	2
y_{max}	2	$\emptyset$	$\emptyset$	2	$\frac{4}{3}$

el máximo de Z es el correspondiente al caso V. Dicho máximo es alcanzado en el punto $x_1 = 1$, $x_2 = 2$, el cual cumple las condiciones de integridad. Por tanto, ésta es la solución del problema siendo el valor de éste $Z = 5$. ♦

4.3. Método del plano de corte

Para evitar el enorme coste que puede llegar a tener el método de ramificación se introduce el método de los planos de corte, el cual optimiza las restricciones que hay que añadir cuando encontramos una solución del problema lineal asociado que no cumple con las condiciones de integridad requeridas de programación lineal entera formulado anteriormente. Sea el problema lineal asociado al problema de programación entera formulado anteriormente dado por

Optimizar $< \overrightarrow{c}, \overrightarrow{x} >$,

sometido a $A\overrightarrow{x} \leq \overrightarrow{b}$, $A \in \mathbb{R}_{m\times n}$.

$x_j \geq 0$, $j \in \{1, \ldots, n\}$.

Para resolver el problema entero supondremos, en primer lugar, que se tiene resuelto el problema lineal asociado y que $\overline{x}$ es una solución óptima del problema lineal asociado, la cual no es entera, pues caso de serlo ésta sería también la solución del problema entero. Un plano de corte es un hiperplano de $\mathbb{R}^n$ que separa estrictamente $\{\overrightarrow{x}\}$ del conjunto de soluciones factibles enteras del problema. Un corte es una restricción lineal de desigualdad que satisface:

1. Elimina la solución óptima del problema lineal asociado que no satisface las condiciones de integridad.
2. Es verificada por todas las soluciones factibles del problema entero inicial.

Si la solución del problema asociado no cumple las condiciones de integridad requeridas, debe darse necesariamente la condición de que reordenadas las ecuaciones de restricción, la matriz D es la correspondiente a la base canónica, la cual suponemos que ocupa las primeras n columnas, y que de entre las variables básicas $x_1, \ldots, x_m$, alguna x_r debe de satisfacer condición de integridad y su respectivo b_r es no entero; entonces para dicha variable x_r se tiene la ecuación:

$$x_r + \sum_{j=m+1}^{n} a_{r,j} = b_r$$

Sea $B_{r,j} = [b_r]$ y $A_{r,j} = [a_{r,j}]$ la parte entera de b_r y $a_{r,j}$ respectivamente, y sean β_r, $\alpha_{r,i}$ las respectivas partes fraccionarias $\beta_r = b_r - B_r$, y $\alpha_{r,j} = a_{r,j} - A_{r,j}$, entonces

$$x_r + \sum_{j=m+1}^{n} [A_{r,j} + \alpha_{r,j}]\, x_{r,j} = B_r + \beta_r$$

Si se reordena esta ecuación como:

$$x_r + \sum_{j=m+1}^{n} a_{r,j}x_j - B_r = \beta_r - \sum_{j=m+1}^{n} \alpha_{i,r}x_j$$

Para cualquier solución factible que satisfaga las condiciones de integridad, el primer miembro será entero. El segundo miembro es pues entero, y puesto que $B_r \in]0,1[$ y $A_{r,j} \in [0,1[$ se tiene que el segundo miembro es también estrictamente menor que la unidad. Así, para cualquier solución factible óptima del problema asociado se verifica la desigualdad:

$$\beta_r - \sum_{j=m+1}^{n} \alpha_{i,r}x_j \leq 0$$

Por otra parte, si $\overrightarrow{y}$ es una solución factible del problema lineal asociado, resultará que puesto que ahora x_j, $j = m+1,\ldots,n$ son variables no básicas (y por tanto, nulas), no podrá verificarse la condición $\beta_r - \sum\limits_{j=m+1}^{n} \alpha_{i,r}x_j = \beta_r > 0$.
Por tanto, el problema entero satisface además de las restricciones iniciales, la restricción adicional deducida del plano de corte.
Así, ahora se procede de nuevo a resolver el problema asociado añadiendo la nueva restricción, procediéndose de nuevo como antes, hasta alcanzar una solución factible del problema asociado con las restricciones añadidas que verifique la condición de integridad.
Nótese que en cada iteración se deberá aplicar, posiblemente, el método de penalización o el método de las dos fases, puesto que la solución óptima del problema asociado no pertenece al conjunto factible en cuanto añadimos la nueva restricción.

Ejemplo 4.3.1

Maximizar la función
$Z = x_1 + 2x_2$
sometida a las restricciones
$4x_1 + 3x_2 \leq 12$
$-x_1 + \quad x_2 \leq \ \ 2$
$x_1, x_2 \geq 0$
$x_1, x_2 \in \mathbb{Z}$

En primer lugar, procedemos a resolver el programa lineal asociado dado por

Maximizar $Z = x_1 + 2x_2$
sometida a las restricciones
$4x_1 + 3x_2 \leq 12$
$-x_1 + \quad x_2 \leq \ \ 2$
$x_1, x_2 \geq 0$

Dicho problema resulta en su forma estándar

$$\begin{array}{l} \text{Maximizar } Z = x_1 + 2x_2 \\ \text{sometida a las restricciones} \\ \;\; 4x_1 + 3x_2 + x_3 \qquad\; = 12 \\ -\;\; x_1 + \;\; x_2 \qquad + x_4 = \;\; 2 \\ x_1, x_2, x_3, x_4 \geq 0 \end{array}$$

problema que ha sido resuelto en el ejemplo 4.2.1, y cuya solución se alcanza en el punto $x_1 = \frac{6}{7}$, $x_2 = \frac{20}{7}$, $x_3 = 0$, $x_4 = 0$.

Dicha solución no cumple el requisito de integridad; por tanto, se buscará un plano de corte para excluir dicha solución de la región factible. Así, se tiene:

$$x_1 + \frac{1}{7}x_3 - \frac{3}{7}x_4 = \frac{6}{7}.$$

Separando cada coeficiente en parte entera y parte fraccionaria, se obtiene

$$x_1 + \frac{1}{7}x_3 - x_4 + \frac{4}{7}x_4 = \frac{6}{7},$$

de donde se tiene

$$x_1 - x_4 = -\frac{1}{7}x_3 - \frac{4}{7}x_4 + \frac{6}{7} \leq 0,$$

por tanto, para obtener integridad debe satisfacerse la relación

$$\frac{1}{7}x_3 + \frac{4}{7}x_4 \geq \frac{6}{7}$$

lo que escrito en forma estándar será

$$\frac{1}{7}x_3 + \frac{4}{7}x_4 - x_5 = \frac{6}{7},$$

donde x_5 es una variable de holgura $x_5 \geq 0$. Puesto que la variable de holgura entra con signo negativo es necesario introducir una variable artificial y aplicar el método de las dos fases; así, en primer lugar, se maximiza la función auxiliar $Z_{aux} = -x_6$ en el conjunto definido por las restricciones dadas en la tabla anterior junto a la nueva dada por el plano de corte. Así, la primera fase en forma tabular se escribe como

Fase I:

Básicas	0	0	0	0	0	-1	0
x_1	1	0	$\frac{1}{7}$	$-\frac{3}{7}$	0	0	$\frac{6}{7}$
x_2	0	1	$\frac{1}{7}$	$\frac{4}{7}$	0	0	$\frac{20}{7}$
x_6	0	0	$\frac{1}{7}$	$\frac{4}{7}$	-1	1	$\frac{6}{7}$

Escribiendo la función objetivo en función de variables no básicas, se tiene

Básicas	0	0	$\frac{1}{7}$	$\frac{4}{7}$	-1	0	$\frac{6}{7}$
x_1	1	0	$\frac{1}{7}$	$-\frac{3}{7}$	0	0	$\frac{6}{7}$
x_2	0	1	$\frac{1}{7}$	$\frac{4}{7}$	0	0	$\frac{20}{7}$
x_6	0	0	$\frac{1}{7}$	$\frac{4}{7}$	-1	1	$\frac{6}{7}$

El mayor coeficiente positivo corresponde a la columna asociada a x_4; así

Básicas	0	0	$\frac{1}{7}$	$\frac{4}{7}$	-1	0	$\frac{6}{7}$
x_1	1	0	$\frac{1}{7}$	$-\frac{3}{7}$	0	0	$\frac{6}{7}$
x_2	0	1	$\frac{1}{7}$	$\frac{4}{7}$	0	0	$\frac{20}{7}$
x_6	0	0	$\frac{1}{4}$	1	-$\frac{7}{4}$	$\frac{7}{4}$	$\frac{3}{2}$

El menor término independiente correspondiente a las filas cuyo valor en la columna de x_4 es 1, corresponde a la tercera fila; así, entrará en la base x_4 y saldrá x_6, por tanto:

Básicas	0	0	0	0	0	-1	0
x_1	1	0	$\frac{1}{4}$	0	$-\frac{3}{4}$	$\frac{3}{4}$	$\frac{3}{2}$
x_2	0	$\frac{7}{4}$	0	0	$\frac{7}{4}$	-$\frac{7}{4}$	$\frac{7}{2}$
x_4	0	0	$\frac{1}{4}$	1	-$\frac{7}{4}$	$\frac{7}{4}$	$\frac{3}{2}$

La función auxiliar se maximiza para $x_6 = 0$ y toma valor $Z_{aux} = 0$

Fase II

Básicas	0	0	$-\frac{3}{7}$	-$\frac{5}{7}$	0	-$\frac{46}{7}$
x_1	1	0	$\frac{1}{4}$	0	$-\frac{3}{4}$	$\frac{3}{2}$
x_2	0	$\frac{7}{4}$	0	0	$\frac{7}{4}$	$\frac{7}{2}$
x_4	0	0	$\frac{1}{4}$	1	-$\frac{7}{4}$	$\frac{3}{2}$

Escribiendo la función objetivo en función de variables no básicas, y escribiendo la base en forma canónica, se tiene

Básicas	0	0	$-\frac{1}{4}$	0	-$\frac{5}{4}$	-$\frac{11}{2}$
x_1	1	0	$\frac{1}{4}$	0	$-\frac{3}{4}$	$\frac{3}{2}$
x_2	0	1	0	0	1	2
x_4	0	0	$\frac{1}{4}$	1	-$\frac{7}{4}$	$\frac{3}{2}$
	$\frac{3}{2}$	2	0	$\frac{3}{2}$	0	

Puesto que todos los coeficientes de la función objetivo, una vez escrita en función de variables no básicas, son menores que cero, se tiene que el punto actual es solución optima del problema. Dicha solución tampoco cumple la condición de integridad; por tanto, será necesario añadir un nuevo plano de corte que excluya dicho punto del conjunto factible. Dicho plano se obtiene a partir de la ecuación correspondiente a la variable básica x_1, cuya solución no es entera:

$$x_1 + \frac{1}{4}x_3 - \frac{3}{4}x_5 = \frac{3}{2},$$

de donde separando partes enteras y fraccionarias se obtiene

$$x_1 + \frac{1}{4}x_3 - x_5 + \frac{1}{4}x_5 = 1 + \frac{1}{2},$$

y por tanto

$$x_1 - x_5 - 1 = -\frac{1}{4}x_3 - \frac{1}{4}x_5 + \frac{1}{2} \leq 0.$$

Así, el nuevo plano de corte resulta

$$\frac{1}{4}x_3 + \frac{1}{4}x_5 \geq \frac{1}{2}.$$

Para resolver el problema con la nueva restricción será necesario aplicar de nuevo el método de las dos fases; así, se tiene

Fase I:

Básicas	0	0	0	0	0	0	-1	0
x_1	1	0	$\frac{1}{4}$	0	$-\frac{3}{4}$	0	0	$\frac{3}{2}$
x_2	0	1	0	0	1	0	0	2
x_4	0	0	$\frac{1}{4}$	1	$-\frac{7}{4}$	0	0	$\frac{3}{2}$
x_7	0	0	$\frac{1}{4}$	0	$\frac{1}{4}$	-1	1	$\frac{1}{2}$

Escribiendo la función objetivo en función de variables no básicas se obtiene

Básicas	0	0	$\frac{1}{4}$	0	$\frac{1}{4}$	-1	0	$\frac{1}{2}$
x_1	1	0	$\frac{1}{4}$	0	$-\frac{3}{4}$	0	0	$\frac{3}{2}$
x_2	0	1	0	0	1	0	0	2
x_4	0	0	$\frac{1}{4}$	1	$-\frac{7}{4}$	0	0	$\frac{3}{2}$
x_7	0	0	$\frac{1}{4}$	0	$\frac{1}{4}$	-1	1	$\frac{1}{2}$

El mayor coeficiente positivo de la función objetivo es el correspondiente a x_3, haciendo la unidad en los valores positivos de la columna, se tiene la tabla:

Básicas	0	0	$\frac{1}{4}$	0	$\frac{1}{4}$	-1	0	$\frac{1}{2}$
x_1	4	0	1	0	-3	0	0	6
x_2	0	1	0	0	1	0	0	2
x_4	0	0	1	4	-7	0	0	6
x_7	0	0	1	0	1	-4	4	2

El menor término independiente corresponde a la cuarta ecuación; así

Básicas	0	0	0	0	0	0	-1	0
x_1	4	0	0	0	-5	4	-4	4
x_2	0	1	0	0	1	0	0	2
x_4	0	0	0	4	-8	4	-4	4
x_3	0	0	1	0	1	-4	4	2

Por tanto, aquí finaliza la fase I.

Fase II

Básicas	0	0	$-\frac{1}{4}$	0	-$\frac{5}{4}$	0	-$\frac{11}{2}$
x_1	4	0	0	0	-4	4	4
x_2	0	1	0	0	1	0	2
x_4	0	0	0	4	-8	4	4
x_3	0	0	1	0	1	-4	2

Escribiendo la función objetivo en función de variables no básicas, y expresando la base en forma canónica, se tiene

Básicas	0	0	0	0	-1	-1	-5
x_1	1	0	0	0	-1	1	1
x_2	0	1	0	0	1	0	2
x_4	0	0	0	1	-2	1	1
x_3	0	0	1	0	1	-4	2

Puesto que todos los coeficientes de la función objetivo, una vez escrita en función de variables no básicas, son negativos; se tiene que el punto actual es la solución óptima del problema, y dado que ésta es entera es también la solución del programa entero.

Por tanto, la función $Z = x_1 + 2x_2$ alcanza su máximo en el punto de coordenadas $x_1 = 1$, $x_2 = 2$, siendo el valor del problema $Z = 5$.

Tema 5

Resolución de ecuaciones

5.1. Introducción

En este capítulo se aborda el estudio de un conjunto de técnicas numéricas para la resolución de ecuaciones, tanto algebraicas como trascendentes, hasta alcanzar la solución con un grado de precisión especificado de antemano.
Son pocas las ecuaciones que podemos encontrar que resulten resolubles de un modo exacto; ello no es impedimento para alcanzar sus soluciones con la precisión que se requiera, para lo que se estudian métodos que proporcionan algoritmos útiles a la hora de abordar tales problemas, los cuales son fácilmente implementables en el ordenador mediante el uso de lenguajes de programación de alto nivel, tales como Fortran 2003 o C 99.
Los métodos, los clasificaremos en cerrados si parten de un intervalo cerrado $[a, b]$ donde se sabe que contiene una solución de la ecuación $f(x) = 0$, y abiertos cuando únicamente se cuenta con una aproximación inicial a la solución. Finalmente, por su importancia, se aborda el estudio de los métodos de punto fijo, así como el estudio de su convergencia.

5.2. Métodos cerrados.

El problema que abordan los métodos cerrados es el siguiente: Sea $f : [a, b] \rightarrow \mathbb{R}$ una función continua de modo que presenta signos contrarios en sus extremos, $f(a) \cdot f(b) < 0$; entonces de acuerdo al teorema de Bolzano $\exists c \in]a, b[$ de modo que $f(c) = 0$. Los métodos cerrados que estudiaremos son el método de bisección y el método de la cuerda (o regula falsi).

5.2.1. Método de bisección

El método de bisección consiste en considerar el punto medio $c = \frac{a+b}{2}$ del intervalo $[a, b]$ donde existe una raíz y estudiar el signo de $f(c)$ pudiendo resultar tres casos:

1. $f(c) = 0$, en cuyo caso c es la raíz buscada y por tanto finaliza el proceso.

2. $f(a) \cdot f(c) > 0$, en cuyo caso el signo de f es el mismo en a y en c por lo que el cambio de signo en la función se produce en el intervalo $[c, b]$.

3. $f(a) \cdot f(c) < 0$, en cuyo caso el cambio de signo de f se produce en el intervalo $[a, c]$.

A continuación, se procede a aplicar de nuevo el método al nuevo intervalo donde se produce el cambio de signo, obteniéndose en cada iteración un problema equivalente en un nuevo intervalo cuya longitud es la mitad.
El error que se produce al aproximar la raíz por el punto central del intervalo es menor o igual a la mitad de la longitud de dicho intervalo, y por tanto, para la iteración n-ésima se puede asegurar que éste es menor que $\frac{b-a}{2^n}$.
El método de bisección se materializa en el siguiente algoritmo:

Dada una función continua $f : [a, b] \to \mathbb{R}$ de modo que se satisface $f(a) \cdot f(b) < 0$, y dada una cota de error admisible $\varepsilon > 0$,

1. Tomar $a_0 = a$, $b_0 = b$, $n = 0$.

2. $n = n + 1$, $c_n = \frac{a_{n-1}+b_{n-1}}{2}$.

3. Si

$$\begin{cases} f(c_n) = 0 & r = c_n \text{ solución exacta y fin del algoritmo.} \\ f(a) \cdot f(c) > 0 & a_n = c_n, \quad b_n = b_{n-1}. \\ f(a) \cdot f(c) < 0 & a_n = a_{n-1}, \quad b_n = c_n. \end{cases}$$

4. Si

$$\begin{cases} b_n - a_n < \varepsilon & \text{tomar } c_n \text{ como raíz y finalizar.} \\ b_n - a_n \geq \varepsilon & \text{ir a paso 2.} \end{cases}$$

de este modo se obtiene la raíz $r = c_n \pm \varepsilon$ con la precisión requerida.

La convergencia del método está garantizada pues de no encontrar la raíz exacta ($f(c_n) = 0$, en cuyo caso $r = c_n$ es el verdadero valor de la raíz), $r \in [a_n, b_n]$, $\forall n \in \mathbb{N}$ y, por tanto, $r \in \bigcap_{n\in\mathbb{N}} [a_n, b_r]$ y puesto que $[a_{n+1}, b_{n+1}] \subset [a_n, b_n]$, $\forall n \in \mathbb{N}$ y la longitud de $[a_n, b_n]$ decrece en progresión geométrica, existe un único punto común a todos los intervalos $[a_n, b_n]$, que es la raíz.

Ejemplo 5.2.1

Encontrar una raíz positiva de la ecuación $x - 2 \operatorname{sen} x = 0$ exacta hasta la segunda cifra decimal aplicando el método de bisección en el intervalo [1.3 , 3.5].

En primer lugar se comprueban las condiciones del método; f debe ser una función continua (lo es) que toma valores de signo contrario en los extremos del intervalo (los toma, pues $f(1.3) = -0,63$, $f(3.5) = 4.20$). El método de bisección se basa en tomar $a = 1.3$, $b = 3.5$ y $c = \frac{a+b}{2}$ y reducir el problema al intervalo

$[a, c]$ o $[c, b]$ eligiéndose aquél en el que produce el cambio de signo en f (si resulta $f(c) = 0$, se habría encontrado la raíz y se finalizaría el algoritmo). El proceso se repite hasta que la longitud del intervalo resultante sea menor que la cota de error admitida, momento en el que se toma c como la raíz de f.

n	a	b	c	$f(c)$
1	1.3	3.5	2.4	1.05
2	1.3	2.4	1.85	-0.07
3	1.85	2.4	2.12	0.42
4	1.85	2.12	1.99	0.16
5	1.85	1.99	1.92	0.04
6	1.85	1.92	1.89	-0.02
7	1.89	1.92	1.90	0.01
8	1.89	1.98	1.90	0.01

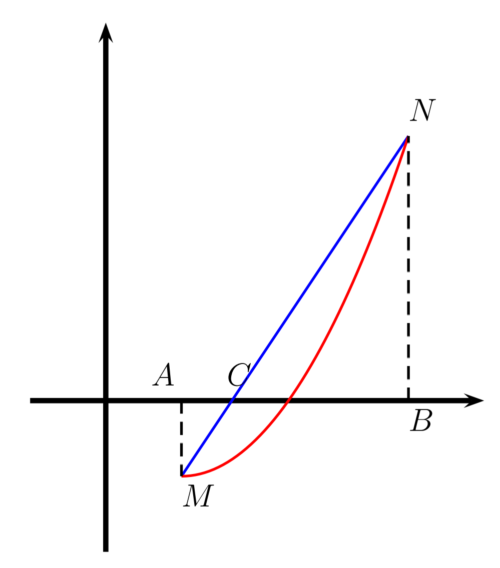
La longitud del intervalo es 0.01, por tanto, la solución es $x = 1.90$.
♦

5.2.2. Método de la regula falsi

El método de la regula falsi, también llamado método de la cuerda, método de las partes proporcionales, o método de la falsa posición, pretende resolver al igual que el método de bisección el problema siguiente:

Dada de una función $f : [a, b] \to \mathbb{R}$ continua tal que $f(a) \cdot f(b) < 0$, encontrar un punto $c \in]a, b[$ de modo que $f(c) = 0$.

Este método en lugar de tomar en cada paso como valor aproximado de la raíz el punto medio del intervalo $[a, b]$, toma un punto c, de modo que los segmentos $\overline{ac}$ y $\overline{cb}$ sean proporcionales a los valores $|f(a)|$ y $|f(b)|$ respectivamente, lo cual proporciona en general mejores aproximaciones.



donde $A = (a, 0)$, $C = (c, 0)$, $B = (b, 0)$, $M = (a, f(a))$, y $N = (b, f(b))$.

En la figura se puede apreciar que los triángulos $\overset{\triangle}{ACM}$ y $\overset{\triangle}{BCN}$ son semejantes, por

tanto, se verifica la igualdad

$$\frac{\overline{AC}}{\overline{BC}} = \frac{\overline{AM}}{\overline{BN}} \quad \text{y por tanto} \quad \frac{c-a}{b-c} = \frac{-f(a)}{f(b)}$$

de donde se obtiene sin dificultad

$$c = \frac{a\,f(b) - b\,f(a)}{f(b) - f(a)}$$

lo que es equivalente a

$$c = a - \frac{b-a}{f(b)-f(a)} f(a) = b - \frac{b-a}{f(b)-f(a)} f(b)$$

Teorema 5.2.1
Sea $f :\to [a,b]$ una función de clase $\mathcal{C}^2([a,b])$ de modo que $f(a)\cdot(b) < 0$ y $\operatorname{sign}(f''(x)) = \text{Cte}\ \forall x \in [a,b]$. Entonces, la sucesión formada por $x_{n+1} = b_n - \frac{f(b_n)}{f(b_n)-f(a_n)}(b_n - a_n)$ donde $a_0 = a$, $b_0 = b$, siendo $a_{n+1} = a_n$, $b_{n+1} = x_{n+1}$ si $f(x_{n+1})\cdot f(a_n) \le 0$, y $a_{n+1} = x_{n+1}$, $b_{n+1} = b_n$ si $f(x_{n+1})\cdot f(a_n) > 0$ converge a una raíz $r \in]a,b[$ de la ecuación $f(x) = 0$.

Dem:

Supongamos que $f''(x) > 0\ \forall x \in [a,b]$ (el caso $f''(x) < 0$ se reduce a éste considerando la función $g(x) = -f(x)$) y supongamos también que $f(a) < 0$ (en otro caso bastará considerar la función $h(x) = f(-x)$ y tomar como intervalo $[-b,-a]$).

Entonces $x_{n+1} = b_n - \frac{f(b_n)}{f(b_n)-f(a_n)}(b_n - a_n)$ y puesto que $f''(x) > 0$ en $[a,b]$ se tiene que la función está por debajo de la cuerda que une los extremos $(a_n, f(a_n))$, $(b_n, f(b_n))$ y, por tanto, $f(x_{n+1}) < 0$. De este modo, $a_n = a$, y $b_{n+1} = x_{n+1}$ $\forall n \in \mathbb{N}$, por tanto, se tiene:

$$x_{n+1} = x_n - \frac{f(b_n)}{f(b_n) - f(a)}(x_n - a)\,\forall n \in \mathbb{N}.$$

Entonces, se tiene que $x_{n+1} - x_n = \frac{-f(x_n)}{f(x_n)-f(a)}(x_n - a) > 0$ por tanto $x_{n+1} < x_n$. Así, la sucesión $\{x_n\}_{n=1}^{\infty}$ es monótona creciente acotada inferiormente por a y, por tanto, $\exists r = \lim\limits_{n\to\infty} x_n$.

Puesto que f es continua, tomando límite en la expresión que define x_{n+1} se tiene

$$\lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} x_n - \frac{\lim\limits_{n\to\infty} x_n}{\lim\limits_{n\to\infty} f(x_n) - f(a)}(\lim_{n\to\infty} x_n - a)$$

y, por tanto:

$$0 = \frac{f(r)}{f(r) - f(a)}(r - a)$$

de donde se tiene que $f(r) = 0$, pues el caso $r = a$ no es posible por ser $f(a) > 0$ y $f(r) = \lim\limits_{n\to\infty} f(x_n) \le 0$. ▼

Para estimar el error cuando el intervalo $[a,b]$ es tal que el signo de $f'(x)$ es constante, puede procederse como: sea r la raíz de $f(x)$ en $[a,b]$ y sea $\{x_n\}_{n=0}^{\infty}$ la sucesión formada por los sucesivos valores de c. Entonces, se verifica:

$$a_n = x_{n-1} - \frac{f(x_{n-1})}{f(x_{n-1}) - f(a)}(x_{n-1} - a)$$

y por tanto, dado que $f(r) = 0$ se tiene

$$f(r) - f(x_{n-1}) = \frac{f(x_{n-1}) - f(a)}{x_{n-1} - a}(x_n - x_{n-1}).$$

Por otra parte:

$$f(x_n) - f(r) = f'(\xi)(x_n - r), \xi \in]a,b[$$
$$f(x_{n-1}) - f(a) = (x_{n-1} - a)f'(\eta)$$

donde $\xi, \eta \in]a,b[$. Así, se tiene

$$(r - x_{n-1})f'(\xi) = (x_n - x_{n-1})f'(\eta)$$

y de aquí

$$|r - x_n| = \frac{|f'(\eta) - f'(\xi)|}{|f'(\xi)|}|x_n - x_{n-1}| \leq \frac{M - m}{m}|x_n - x_{n-1}|$$

donde m y M son el máximo y el mínimo de $|f'(x)|$ en $[a,b]$. Si el intervalo $[a,b]$ es tan pequeño que $M \leq 2m$ se tiene que

$$|r - x_n| \leq |x_n - x_{n-1}|$$

puede tomarse como cota del error.

Ejemplo 5.2.2
Encontrar una raíz positiva de la ecuación $x - 2\operatorname{sen} x = 0$ exacta hasta la segunda cifra decimal (úsese calculadora) aplicando el método de la regula falsi en $[1.3, 3.5]$

En este caso, el valor de c viene dado por

$$c = a - f(a)\frac{b - a}{f(b) - f(a)}$$

deteniéndose el mismo, cuando obtenemos dos valores de c consecutivos cuya diferencia es menor que el error admitido.

n	a	b	c	$f(c)$	$\lvert c_n - c_{n-1}\rvert$
1	1.3	3.5	1.59	-0.41	
2	1.59	3.5	1.75	-0.20	.14
3	1.75	3.5	1.84	-0.09	.09
4	1.84	3.5	1.87	-0.03	.03
5	1.87	3.5	1.89	-0.01	.02
6	1.89	3.5	1.89	0.00	.00

La solución tomada será $x = 1.89$.

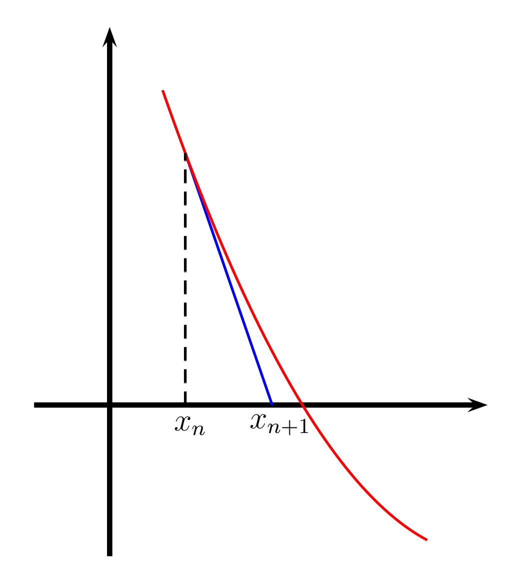
♦

5.3. Métodos abiertos

Los métodos abiertos parten de una aproximación inicial a la raíz x_0, la cual se pretende refinar en sucesivas iteraciones. Los principales métodos abiertos son el método de Newton (y sus modificaciones) y los métodos iterativos de punto fijo.

5.3.1. Método de Newton

A continuación presentamos el llamado método de Newton (también llamado método de la tangente), el cual consiste en dada una ecuación $f(x) = 0$ donde f es una función continua y derivable y dada una aproximación inicial x_0 a una raíz, se construye una nueva aproximación x_1 como el punto de corte de la tangente a $f(x)$ en el punto x_0 con el eje de abscisas, y así sucesivamente.



Sea x_n el punto actual, la ecuación de la recta tangente a la curva $y = f(x)$ en el punto $(x_n, f(x_n))$ viene dada en su forma punto pendiente por

$$y - f(x_n) = f'(x_n)(x - x_n).$$

Esta recta corta al eje de abscisas en el punto $(x_{n+1}, 0)$. Así, se tiene la fórmula de iteración de Newton

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

El error en el método de Newton también se puede estimar mediante

$$|r - x_n| \leq |x_n - x_{n-1}|.$$

Ejemplo 5.3.1
Encontrar una raíz positiva de la ecuación $x - 2\operatorname{sen} x = 0$ exacta hasta la segunda cifra decimal (úsese calculadora) aplicando el método de Newton partiendo de $x_0 = 1.3$.

El método de Newton se basa en la sucesión $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$, deteniéndose el proceso cuando $|x_{n+1} - x_n| < 10^{-2}$. Así, tenemos:

$$f(x) = x - \operatorname{sen} x, \qquad f'(x) = 1 - 2\cos\, x$$

y por tanto,

$$x_{n+1} = x_n - \frac{x_n - 2\operatorname{sen} x_n}{1 - 2\cos\, x_n}$$

Tomando como $x_0 = 1.3$ se obtiene

x_0	x_1	x_2	x_3	x_4
1.30	2.65	2.03	1.90	1.90

por tanto, la solución es $x = 1.90$. ♦

5.3.2. Método de Newton modificado

Consiste en aproximar en todas las iteraciones el valor de $f'(x_n)$ por $f'(x_0)$, de este modo, nos evitaremos tener que calcular la derivada en cada etapa, pero presenta el inconveniente de empeorar la convergencia. Así, se tiene la aproximación

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)}.$$

El error en este método puede estimarse como $|r - x_{n+1}| \approx |x_{n+1} - x_n|$.

5.3.3. Método de la secante

El método de la secante puede considerarse como una variación del método de Newton $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$, pues resulta de sustituir la derivada $f'(x_n)$ por una aproximación a la misma: $f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$ resultando de este modo

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n)$$

El error en este método puede estimarse como $|r - x_{n+1}| \approx |x_{n+1} - x_n|$.

Ejemplo 5.3.2
Explicar el comportamiento de los métodos de bisección, de la regula falsi y de Newton en el caso de que la ecuación a resolver fuera $x^2 - 4x\operatorname{sen} x + 4\operatorname{sen}^2 x = 0$ tomando en caso de los primeros, el intervalo inicial $[1.3, 3.5]$; y $x = 1.3$, en el método de Newton.

Respecto al comportamiento de los métodos para la ecuación $x^2 - 4\,x\operatorname{sen} x + 4\operatorname{sen}^2 x = 0$, se tiene que $f(x)$ es una función continua en $[1.3, 3.5]$ siendo $f(1.3) = 0.39$ y $f(3.5) = 17.65$, dado que ambos valores tienen el mismo signo, no es posible aplicar ni el método de bisección ni el método de la regula falsi.

El método de Newton es un método abierto, por lo que sí que es posible utilizarlo. Así, se tiene:

$f(x) = x^2 - 4x\text{sen}\, x + 4\text{sen}^2\, x, \qquad f'(x) = 2x - 4\text{sen}\, x - 4x\cos\, x + 8\text{sen}\, x \cos\, x$

y así

$$x_{n+1} = x_n - \frac{x_n^2 - 4x_n \text{sen}\, x_n + 4\,\text{sen}^2 x_n}{2x_2 - 4\,\text{sen}\, x_n - 4,\; x_n \cos x_n + 8\,\text{sen}\, x_n \cos x_n}$$

Tomando como $x_0 = 1.3$ se obtiene la sucesión:

x_0	x_1	x_2	x_3	x_4
1.30	1.97	1.93	1.90	1.90

por tanto, la solución es $x = 1.90$.

♦

5.3.4. Métodos de punto fijo

A continuación se estudia un tipo particular de funciones, las funciones contractivas, útiles para la construcción de métodos numéricos de resolución de ecuaciones. Estos métodos están basados en el teorema del punto fijo que estudiamos a continuación.

Definición 5.3.1
Sea $f : I \to I$ una aplicación de un intervalo I, acotado o no, de $\mathbb{R}$. Diremos que f es una **aplicación contractiva** si $\exists k \in \mathbb{R}$, $0 < k < 1$ de modo que $\forall x_1, x_2 \in I$ se verifica $|f(x_1) - f(x_2)| \leq K\, |x_1 - x_2|$ (el caso de $K = 0$ no se considera pues implica f constante en I).

Teorema 5.3.1
Si $f : I \to I$ es contractiva, entonces f es continua en I.

Dem:

Sea $\epsilon > 0$, entonces se tiene que si $x_1, x_2 \in I$ tales que $|x_1 - x_2|\, \delta$ con $\delta = \epsilon$ se verifica

$$|f(x_1) - f(x_2)| \leq K\, |x_1 - x_2| \leq K\delta = K\epsilon < \epsilon$$

por tanto, f es continua en I.

▼

Teorema 5.3.2
Sea $f : I \to \mathbb{R}$ una función continua y derivable en I de modo que $|f'(x)| \leq K < 1$, $\forall x \in I$, entonces f es contractiva en I.

Dem:

Sean $x_1, x_2 \in I$. Por el teorema del valor medio existe $\xi \in]x_1, x_2[$ (suponemos sin pérdida de generalidad que $x_1 < x_2$) tal que $f(x_2) - f(x_1) = f'(\xi)(x_2 - x_1)$ y por tanto:

$$|f(x_2) - f(x_1)| = |f'(\xi)|\, |x_2 - x_1| < K\, |x_2 - x_1|$$

con $0 < K < 1$ y por tanto, f es contractiva.

▼

Definición 5.3.2
Sea $f : A \to A$ una aplicación de un conjunto A en sí mismo, Diremos que un punto $c \in A$ es **punto fijo** de la aplicación f si se verifica que $f(c) = c$.

Teorema 5.3.3 (del punto fijo)
Sea $f : I \to I$ una aplicación contractiva sobre un intervalo de $\mathbb{R}$. Entonces existe un único punto $c \in I$ fijo por la aplicación f.

Dem:
Veamos en primer lugar que si existe, tal punto c es único. Sean $c, r \in I$ puntos fijos de f, entonces, si $c \neq r$ se verifica

$$0 < |c - r| = |f(c) - f(r)| \leq K\,|c - r| \langle\, |c - r|$$

lo cual es absurdo; por tanto, $c = r$.
A continuación veamos la existencia de tal punto fijo. Sea $x_0 \in I$, sea $x_{n+1} = f(x_n)$. Veamos que la sucesión $\{x_n\}_{n=0}^{\infty}$ es una sucesión de Cauchy; para ello, sean $n, m \in \mathbb{N}$ con $m > n$ $m = n + p$. Así, se tiene

$$|x_m - x_n| = |x_{n+p} - x_{n+p-1} + x_{n+p-1} - x_{n+p-2} - \cdot\cdot + x_{n+2} - x_{n+1} + x_{n+1} - x_n| \leq$$

$$\leq |x_{n+p} - x_{n+p-1}| + |x_{n+p-1} - x_{n+p-2}| + \cdots + |x_{n+1} - x_n| \leq \sum_{p=0}^{\infty} |x_{n+p+1} - x_{n+p}|\,.$$

Por otra parte, se verifica $|x_2 - x_1| = |f(x_1) - f(x_0)| \leq K\,|x_1 - x_0|$, y por tanto, por inducción se tiene $|x_{p+1} - x_p| \leq K^p\,|x_1 - x_0|$; por tanto, se tiene

$$|x_m - x_n| \leq K^{n+p-1}\,|x_1 - x_0| + K^{n+p-2}\,|x_1 - x_0| + \cdot\cdot + K^n\,|x_1 - x_0| \leq$$

$$\leq \sum_{p=0}^{\infty} K^{n+p}\,|x_1 - x_0| = \frac{K^n}{1-K}\,|x_1 - x_0|\,.$$

Por tanto, para que $|x_m - x_n| < \epsilon$ es suficiente que $\frac{K^n}{1-K}\,|x_1 - x_0| < \epsilon$, lo cual en el caso de que $x_0 = x_1$ (y, por tanto, x_0 punto fijo de f) se cumple, y en el caso de ser $x_0 \neq x_1$ bastará que

$$K^n < (1-K)\epsilon, \quad n \log K < \log\left[(1-K)\epsilon\right], \quad n > \frac{\log\left[(1-K)\epsilon\right]}{\log K}$$

pues $\log K < 0$ al ser $K \in]0, 1[$, por tanto, para que $|x_m - x_n| < \epsilon$ bastará que $n, m \geq E\left[\frac{\log[(1-K)\epsilon]}{\log K}\right] + 1$ siendo E la función parte entera.
La sucesión $\{x_n\}_{n=0}^{\infty}$ es de Cauchy y, por tanto, es convergente; sea $x = \lim\limits_{n\to\infty} x_n$, por continuidad se tiene que $x = \lim\limits_{n\to\infty} x_{n+1} = \lim\limits_{n\to\infty} f(x_n) = f(\lim\limits_{n\to\infty} x_n) = f(x)$ y por tanto x es punto fijo de f.

▼

Dada una ecuación $f(x) = 0$, puede en general escribirse de una forma equivalente como $x = \phi(x)$, si ϕ es una aplicación contractiva sobre un intervalo I, puede obtenerse la solución de la ecuación inicial tomando un punto $x_0 \in I$ y a partir de el construyendo la sucesión $x_{n+1} = \phi(x_n)$; así, se tendrá que $x = \lim\limits_{n\to\infty} x_n$ es la raíz buscada.

Ejemplo 5.3.3
Mediante el método iterativo resolver la ecuación $x + 0.2\,\mathrm{sen}\,x + 0.1\cos x = 5$ partiendo de la aproximación $x_0 = 5$. ¿Qué se puede decir acerca de la convergencia del método?

Para resolver el problema mediante el método iterativo (lo haremos hasta obtener cuatro cifras decimales exactas), construimos en primer lugar la relación de recurrencia

$$x_{n+1} = 5 - 0.2\mathrm{sen}\,x_n - 0.1\cos\,x_n, \qquad x_0 = 5$$

lo que conduce a la tabla:

n	0	1	2	3	4	5	6	7
x_n	5	5.16342	5.13641	5.14115	5.14032	5.14047	5.14044	5.14045

por tanto, la solución con cuatro cifras decimales exactas es $x = 5.1405$.

Respecto a la convergencia del método, observar que la relación de recurrencia está obtenida a partir de la función $\phi(x) = 5 - 0.2\mathrm{sen}\,x - 0.1\cos\,x$ función continua y derivable en toda la recta real. Esta función satisface

$$|\phi'(x)| = |-0.2\cos\,x + 0.1\mathrm{sen}\,x| \leq |0.2\cos\,x| + |0.1\mathrm{sen}\,x| =$$

$$= 0.2|\cos\,x| + 0.1|\mathrm{sen}\,x| \leq 0.2 + 0.1 = 0.3 < 1$$

por tanto, el método converge.

♦

5.4. Sistemas de ecuaciones no lineales

A continuación se aborda el estudio de la resolución de un sistema de ecuaciones de la forma

$$\begin{array}{c} f_1(x_1, x_2, ..., x_n) = 0 \\ f_1(x_1, x_2, ..., x_n) = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ f_1(x_1, x_2, ..., x_n) = 0 \end{array}$$

el cual, en forma vectorial se escribe como $\overrightarrow{f}(\overrightarrow{x}) = \overrightarrow{0}$, y del que tenemos una aproximación inicial $\overrightarrow{x}_0 = (x_1^0, x_2^0, ..., x_n^0)$.

A continuación estudiaremos métodos para construir una sucesión de aproximaciones convergente a la solución.

5.4.1. Método de Newton-Ralphson

El método de Newton-Ralphson consiste en dada una aproximación $\overrightarrow{x}_k$ a la solución de la ecuación $\overrightarrow{f}(\overrightarrow{x}) = \overrightarrow{0}$, buscar una nueva iteración $\overrightarrow{x}_{k+1} = \overrightarrow{x}_k +$

$\Delta\overrightarrow{x}_k$ de modo que satisfaga no la ecuación inicial $\overrightarrow{f}(\overrightarrow{x}) = \overrightarrow{0}$, sino la ecuación linealizada en $\Delta\overrightarrow{x}_k$, esto es el desarrollo de Taylor en primer orden en $\Delta\overrightarrow{x}_k$ de $\overrightarrow{f}(\overrightarrow{x}_{k+1})$ alrededor del punto $\overrightarrow{x_k}$. Dicha ecuación se escribe en forma vectorial como

$$\overrightarrow{f}(\overrightarrow{x}_{k+1}) = \overrightarrow{f}(\overrightarrow{x}_k + \Delta\overrightarrow{x}_k) \approx \overrightarrow{f}(\overrightarrow{x}_k) + \mathcal{J}(\overrightarrow{f})(\overrightarrow{x}_k)\cdot\Delta\overrightarrow{x}_k = \overrightarrow{0}$$

donde $\mathcal{J}$ denota la matriz jacobiana de la función $\overrightarrow{f}$.
Si $\mathcal{J}(\overrightarrow{f})(\overrightarrow{x}_k) \neq 0$ la ecuación

$$\overrightarrow{f}(\overrightarrow{x}_k) + \mathcal{J}(\overrightarrow{f})(\overrightarrow{x}_k)\cdot\Delta\overrightarrow{x}_k = \overrightarrow{0}$$

puede multiplicarse por $\left[\mathcal{J}(\overrightarrow{f})(\overrightarrow{x}_k)\right]^{-1}$, obteniéndose

$$\left[\mathcal{J}(\overrightarrow{f})(\overrightarrow{x}_k)\right]^{-1}\cdot\overrightarrow{f}(\overrightarrow{x}_k) + \Delta\overrightarrow{x}_k = \overrightarrow{0}$$

de donde

$$\Delta\overrightarrow{x}_k = -\left[\mathcal{J}(\overrightarrow{f})(\overrightarrow{x}_k)\right]^{-1}\cdot\overrightarrow{f}(\overrightarrow{x}_k), \qquad \overrightarrow{x}_{k+1} = \overrightarrow{x}_k + \Delta\overrightarrow{x}_k$$

lo que constituye la fórmula de iteración de Newton-Ralphson.

La fórmula anterior requiere la inversión de una matriz, lo cual constituye una de las operaciones mas difíciles en cálculo numérico, por ello es preferible, en general, la resolución del sistema $J(\overrightarrow{f})(\overrightarrow{x}_k\cdot)\Delta\overrightarrow{x}_k = -\overrightarrow{f}(\overrightarrow{x}_k)$, el cual resulta de modo explícito:

$$\begin{bmatrix} \left.\frac{\partial f_1}{\partial x_1}\right|_{\overrightarrow{x}_k} & \left.\frac{\partial f_1}{\partial x_2}\right|_{\overrightarrow{x}_k} & \dots & \left.\frac{\partial f_1}{\partial x_n}\right|_{\overrightarrow{x}_k} \\ \left.\frac{\partial f_2}{\partial x_1}\right|_{\overrightarrow{x}_k} & \left.\frac{\partial f_2}{\partial x_2}\right|_{\overrightarrow{x}_k} & \dots & \left.\frac{\partial f_2}{\partial x_n}\right|_{\overrightarrow{x}_k} \\ \dots & \dots & \dots & \dots \\ \left.\frac{\partial f_n}{\partial x_1}\right|_{\overrightarrow{x}_k} & \left.\frac{\partial f_n}{\partial x_2}\right|_{\overrightarrow{x}_k} & \dots & \left.\frac{\partial f_n}{\partial x_n}\right|_{\overrightarrow{x}_k} \end{bmatrix} \begin{bmatrix} \Delta x_1^k \\ \Delta x_2^k \\ \dots \\ \Delta x_n^k \end{bmatrix} = \begin{bmatrix} -f_1(\overrightarrow{x}_k) \\ -f_2(\overrightarrow{x}_k) \\ \dots \\ -f_n(\overrightarrow{x}_k) \end{bmatrix}$$

Teorema 5.4.1
Sea un sistema de ecuaciones lineales dado por

$$\begin{gathered} f_1(x_1, x_2, ..., x_n) = 0 \\ f_1(x_1, x_2, ..., x_n) = 0 \\ \dots\dots\dots\dots\dots \\ f_1(x_1, x_2, ..., x_n) = 0 \end{gathered}$$

donde $f \in \mathcal{C}^2(\Omega)$, donde $\Omega \subset \mathbb{R}^n)$ es un conjunto abierto y convexo. Sea $\overrightarrow{x}_0 \in \Omega$ tal que $\exists\varepsilon \geq 0$ para el cual $V = \{\overrightarrow{x} \in \mathbb{R}^n : ||\overrightarrow{x} - \overrightarrow{x}_0|| \leq \epsilon\} \subset \Omega$ siendo $||\overrightarrow{x}|| = \text{máx}\{|x_1| : i = 1, \dots, n\}$, de manera que se satisfagan las condiciones

- La matriz jacobiana $\mathcal{J}\overrightarrow{f}(\overrightarrow{x}_0)$ tiene inversa $\Gamma_0 = \mathcal{J}\overrightarrow{f}^{-1}(\overrightarrow{x}_0)$ de modo que $||\Gamma_0|| \leq \delta$.

- $||\Gamma_0 \overrightarrow{f}(\overrightarrow{x}_0)|| \leq \frac{\varepsilon}{2}$.
- $\sum_{k=1}^{n} \left| \frac{\partial^2 f_i}{\partial x_j \partial x_k} \right| \leq \eta, \ \forall i, j = 1, n \ \overrightarrow{x} \in V$.
- $n\epsilon\delta\eta \leq 1$.

Entonces la sucesión

$$\overrightarrow{x}_{n+1} = \overrightarrow{x}_n - \left[\mathcal{J}(\overrightarrow{f})(\overrightarrow{x}_n)\right]^{-1} \cdot \overrightarrow{f}(\overrightarrow{x}_n)$$

es convergente a un vector $\overrightarrow{x}$ de modo que $\overrightarrow{f}(\overrightarrow{x}) = 0$.

Dem:

No se proporciona por exceder el nivel del curso. ▼

Ejemplo 5.4.1

Resolver mediante el método de Newton-Ralphson el sistema:

$$1 + x^2 - y^2 + e^x \cos x = 0$$
$$2xy + e^x \operatorname{sen} y = 0$$

partiendo de la aproximación inicial $(x_0, y_0) = (-1, 4)$. Aplicar cinco iteraciones.

El método de Newton-Ralphson consiste, como es sabido, en reemplazar el sistema de ecuaciones

$$f(x, y) = 0$$
$$g(x, y) = 0$$

en el entorno de una aproximación a la solución (x_n, y_n) por su polinomio de Taylor de primer orden; así si $(x_n+\Delta x_n, y_n+\Delta y_n)$ es la solución en primer orden en Taylor de la ecuación inicial:

$$0 = f(x_n + \Delta x_n) \approx f(x_n, y_n) + \left.\frac{\partial f}{\partial x}\right|_{(x_n,y_n)} \Delta x_n + \left.\frac{\partial f}{\partial y}\right|_{(x_n,y_n)} \Delta y_n$$
$$0 = g(x_n + \Delta x_n) \approx g(x_n, y_n) + \left.\frac{\partial g}{\partial x}\right|_{(x_n,y_n)} \Delta x_n + \left.\frac{\partial g}{\partial y}\right|_{(x_n,y_n)} \Delta y_n$$

Las derivadas parciales vienen dadas por:

$$\left.\frac{\partial f}{\partial x}\right|_{(x_n,y_n)} = 2x_n + e^{x_n} \cos\, x_n - e^{x_n} \cos\, x_n, \qquad \left.\frac{\partial f}{\partial y}\right|_{(x_n,y_n)} = -2y_n$$
$$\left.\frac{\partial g}{\partial x}\right|_{(x_n,y_n)} = 2Y_n + e^{x_n} \operatorname{sen} y_n, \qquad \left.\frac{\partial g}{\partial y}\right|_{(x_n,y_n)} = 2x_n + e^{x_n} \cos\, y_n$$

así, para la primera iteración se tiene el sistema:

$$\begin{bmatrix} -1.49167 & -8.00000 \\ 7.72159 & -2.24046 \end{bmatrix} \begin{bmatrix} \Delta x_0 \\ \Delta y_0 \end{bmatrix} = \begin{bmatrix} 13.8012 \\ 8.27841 \end{bmatrix}, \qquad \begin{bmatrix} \Delta x_0 \\ \Delta y_0 \end{bmatrix} = \begin{bmatrix} 0.542214 \\ -1.82626 \end{bmatrix}$$

así, calculamos $x_1 = x_0 + \Delta x_0, y_1 = y_0 + \Delta y_0$ resultando $(x_1, y_1) = (-0.457786, 2.17374)$. Para la segunda iteración se tiene el sistema:

$$\begin{bmatrix} -0.06841 & -4.34749 \\ 4.86861 & -1.27435 \end{bmatrix} \begin{bmatrix} \Delta x_1 \\ \Delta y_1 \end{bmatrix} = \begin{bmatrix} 2.94806 \\ 1.46910 \end{bmatrix}, \qquad \begin{bmatrix} \Delta x_1 \\ \Delta y_1 \end{bmatrix} = \begin{bmatrix} 0.123746 \\ -0.68005 \end{bmatrix}$$

de donde $(x_2, y_2) = (-0.457786, 2.17374)$. Para la tercera iteración se tiene:

$$\begin{bmatrix} 0.24313 & -2.98738 \\ 3.70128 & 0.61292 \end{bmatrix} \begin{bmatrix} \Delta x_2 \\ \Delta y_2 \end{bmatrix} = \begin{bmatrix} 0.44308 \\ 0.284005 \end{bmatrix}, \qquad \begin{bmatrix} \Delta x_2 \\ \Delta y_2 \end{bmatrix} = \begin{bmatrix} 0.05288 \\ -0.14401 \end{bmatrix}$$

de donde $(x_3, y_3) = (-0.28116, 1.34968)$. Para la cuarta iteración se tiene:

$$\begin{bmatrix} 0.37242 & -2.69935 \\ 3.43588 & -0.39674 \end{bmatrix} \begin{bmatrix} \Delta x_3 \\ \Delta y_3 \end{bmatrix} = \begin{bmatrix} 0.01730 \\ 0.02241 \end{bmatrix}, \qquad \begin{bmatrix} \Delta x_3 \\ \Delta y_3 \end{bmatrix} = \begin{bmatrix} 0.00588 \\ -0.00560 \end{bmatrix}$$

de donde $(x_4, y_4) = (-0.27528, 1.34407)$. Finalmente para la quinta iteración se tiene:

$$\begin{bmatrix} 0.38661 & -2.68815 \\ 3.42808 & -0.37987 \end{bmatrix} \begin{bmatrix} \Delta x_4 \\ \Delta y_4 \end{bmatrix} = \begin{bmatrix} -0.00001 \\ 0.00007 \end{bmatrix}, \qquad \begin{bmatrix} \Delta x_4 \\ \Delta y_4 \end{bmatrix} = \begin{bmatrix} 0.00002 \\ 0.00001 \end{bmatrix}$$

de donde $(x_5, y_5) = (-0.27530, 1.34408)$.

♦

5.4.2. Método de la máxima pendiente

Una alternativa para la resolución de sistemas no lineales es la siguiente:
Sea el sistema de ecuaciones lineales

$$\begin{array}{l} f_1(x_1, x_2, ..., x_n) = 0 \\ f_1(x_1, x_2, ..., x_n) = 0 \\ \dots\dots\dots\dots\dots\dots \\ f_1(x_1, x_2, ..., x_n) = 0 \end{array}$$

y sea $(x_1^0, x_2^0, \ldots, x_n^0)$ una aproximación inicial.

El método de la máxima pendiente consiste en transformar el problema de resolver el sistema en el equivalente consistente en encontrar los puntos $\overrightarrow{x} \in \mathbb{R}^n$ de modo que la función potencial $U : \mathbb{R}^n \to \mathbb{R}$ definida como $U(\overrightarrow{x}) = \sum_{i=1}^{n} f_i(\overrightarrow{x})^2$ se anule; $U(\overrightarrow{x}) = 0$.

Sea $\overrightarrow{x}$ un punto en el que se anula $U(\overrightarrow{x})$, entonces dicho punto es un mínimo de $U(\overrightarrow{x})$ y, por tanto, si $\overrightarrow{f} : \mathbb{R}^n \to \mathbb{R}^n$ es de clase $\mathcal{C}^1(\mathbb{R}^n)$ debe verificarse que

$\mathcal{J}\overrightarrow{f}\Big|_{\overrightarrow{x}} = 0$. Por otra parte, indicar que la condición de mínimo de $U(\overrightarrow{x})$ no es suficiente para que $\overrightarrow{x}$ sea raíz de la ecuación $\overrightarrow{f}(\overrightarrow{x}) = 0$, pero sí necesaria.

El método de la máxima pendiente consiste en lo siguiente: partiendo de una aproximación inicial $\overrightarrow{x}_0$ a una raíz de la ecuación $\overrightarrow{f}(\overrightarrow{x}) = 0$, buscar el mínimo de U al que se llega siguiendo la dirección de máxima pendiente. Si además en dicho mínimo se anula la función potencial U, éste será raíz de la ecuación.

Para encontrar el mínimo de U siguiendo la dirección de máximo descenso en $\overrightarrow{x}_k$ más próximo a $\overrightarrow{x}_k$, al cual denotaremos como $\overrightarrow{x}_{k+1}$, se define la función $\Psi_k(t) = U(\overrightarrow{x}_k - t\overrightarrow{\nabla}U(\overrightarrow{x}_k))$.

Si $\overrightarrow{x}_k$ es próximo al mínimo, t será una cantidad pequeña, por ello para evaluar $\Psi(t) = U(\overrightarrow{x}_k - t\nabla U(\overrightarrow{x}_k))$ procederemos a aproximar dicha cantidad por

$$\Psi_k(t) = U(\overrightarrow{x}_k - t\overrightarrow{\nabla U}(\overrightarrow{x}_k)) = \sum_{i=1}^{n} f_i^2(\overrightarrow{x}_k - t\overrightarrow{\nabla}U(\overrightarrow{x}_k)) =$$

$$= \sum_{i=1}^{n} \Big[f_i(\overrightarrow{x}_k) - t\langle\overrightarrow{\nabla}f_i(\overrightarrow{x}_k), U(\overrightarrow{x}_k))\rangle + O(t^2)\Big]^2$$

despreciando términos en t de orden superior al primero, se tiene que la condición de mínimo viene dada por

$$\Psi_k(t) = \sum_{i=1}^{n} \Big[f_i(\overrightarrow{x}_k) - t\langle\overrightarrow{\nabla}f_i(\overrightarrow{x}_k), \overrightarrow{\nabla}U(\overrightarrow{x}_k))\rangle\Big]^2$$

de donde:

$$\frac{d\Psi_k(t)}{dt} = -2\sum_{i=1}^{n} \Big[f_i(\overrightarrow{x}_k) - t\langle\overrightarrow{\nabla}f_i(\overrightarrow{x}_k), \overrightarrow{\nabla}U(\overrightarrow{x}_k))\rangle\Big]\langle\overrightarrow{\nabla}f_i(\overrightarrow{x}_k), \overrightarrow{\nabla}U(\overrightarrow{x}_k))\rangle$$

y por tanto,

$$\frac{d\Psi_k(t)}{dt} = -2\sum_{i=1}^{n} \Big\{f_i(\overrightarrow{x}_k)\langle\overrightarrow{\nabla}f_i(\overrightarrow{x}_k), \overrightarrow{\nabla}U(\overrightarrow{x}_k))\rangle - t\langle\overrightarrow{\nabla}f_i(\overrightarrow{x}_k), \overrightarrow{\nabla}U(\overrightarrow{x}_k))\rangle^2\Big\}$$

Dado que $\mathcal{J}\overrightarrow{f}(\overrightarrow{x})$ es la matriz jacobiana de la función $\overrightarrow{f}$, se tiene que

$$\sum_{i=1}^{n} \Big\{f_i(\overrightarrow{x}_k)\langle\overrightarrow{\nabla}f_i(\overrightarrow{x}_k), U(\overrightarrow{x}_k))\rangle\Big\} = \langle\overrightarrow{f}(\overrightarrow{x}_k), \mathcal{J}\overrightarrow{f}(\overrightarrow{x}_k)\cdot\overrightarrow{\nabla}U(\overrightarrow{x}_k\rangle,$$

y

$$\sum_{i=1}^{n}\langle\nabla f_i(\overrightarrow{x}_k), \nabla U(\overrightarrow{x}_k))\rangle^2 = \langle\mathcal{J}\overrightarrow{f}(\overrightarrow{x}_k)\cdot\overrightarrow{\nabla}U(\overrightarrow{x}_k), \mathcal{J}\overrightarrow{f}(\overrightarrow{x}_k)\cdot\overrightarrow{\nabla}U(\overrightarrow{x}_k)\rangle,$$

y por tanto,

$$\frac{d\Psi_k(t)}{dt} = -2\Big[\langle\overrightarrow{f}(\overrightarrow{x}_k), \mathcal{J}\overrightarrow{f}(\overrightarrow{x}_k)\cdot\overrightarrow{\nabla}U(\overrightarrow{x}_k)\rangle -$$

$$- t\langle\mathcal{J}\overrightarrow{f}(\overrightarrow{x}_k)\cdot\overrightarrow{\nabla}U(\overrightarrow{x}_k), \mathcal{J}\overrightarrow{f}(\overrightarrow{x}_k)\cdot\overrightarrow{\nabla}U(\overrightarrow{x}_k)\rangle\Big],$$

Sea t_k el valor de t que anula la expresión anterior, dicho valor viene dado por

$$t_k = \frac{\langle \overrightarrow{f}(\overrightarrow{x}_k), \mathcal{J}\overrightarrow{f}(\overrightarrow{x}_k)\cdot\overrightarrow{\nabla}U(\overrightarrow{x}_k)\rangle}{\langle \mathcal{J}\overrightarrow{f}(\overrightarrow{x}_k)\cdot\overrightarrow{\nabla}U(\overrightarrow{x}_k), \mathcal{J}\overrightarrow{f}(\overrightarrow{x}_k)\cdot\overrightarrow{\nabla}U(\overrightarrow{x}_k)\rangle}.$$

de donde:

$$\overrightarrow{x}_{k+1} = \overrightarrow{x}_k - \frac{\langle \overrightarrow{f}(\overrightarrow{x}_k), \mathcal{J}\overrightarrow{f}(\overrightarrow{x}_k)\cdot\overrightarrow{\nabla}U(\overrightarrow{x}_k)\rangle}{\langle \mathcal{J}\overrightarrow{f}(\overrightarrow{x}_k)\cdot\overrightarrow{\nabla}U(\overrightarrow{x}_k), \mathcal{J}\overrightarrow{f}(\overrightarrow{x}_k)\cdot\overrightarrow{\nabla}U(\overrightarrow{x}_k)\rangle}\overrightarrow{\nabla}U(\overrightarrow{x}_k)$$

Para ello puede procederse según el siguiente algoritmo:

1. Determinar la dirección de máximo decrecimiento en el punto $\overrightarrow{x}_k$, la cual viene dada por $-\nabla U(\overrightarrow{x}_k)$ y el jacobiano de $\overrightarrow{f}(\overrightarrow{x}_k)$ dado por $\mathcal{J}(\overrightarrow{f})(\overrightarrow{x}_k)$.
2. Buscar el mínimo de $U(\overrightarrow{x})$ siguiendo la dirección de máximo descenso $\overrightarrow{x}_{k+1}$ dado por
3. Evaluar la condición de error de modo que si $||\overrightarrow{x}_{k+1} - \overrightarrow{x}_k|| \leq \varepsilon$ se toma como aproximación $\overrightarrow{x}_{k+1}$ y finaliza el algoritmo. Caso de no cumplirse la condición, ir al paso 2.

5.5. Ecuaciones polinómicas

En esta sección se aborda de modo especial el estudio de las ecuaciones de tipo polinómico, esto es, ecuaciones de la forma

$$a_n z^n + a_{n-1}z^{n-1} + \ldots + a_1 z + a_0 = 0,\ a_0, a_1, \ldots, a_n \in \mathbb{C},\ a_n \neq 0,\ z \in \mathbb{C}$$

donde $\mathbb{C}$ representa el cuerpo de los números complejos.

Definición 5.5.1
Sea $P_n(z) = a_n z^n + a_{n-1}z^{n-1} + \ldots + a_1 z + a_0$ un polinomio de grado $n \geq 1$. Diremos que $r \in \mathbb{C}$ es **raíz** de $P_n(z)$ si se verifica que $P_n(r) = 0$.

Algunos resultados clásicos sobre polinomios son los siguientes:

Teorema 5.5.1 (Teorema fundamental del álgebra)
Sea $P_n(z) = \sum\limits_{i=0}^{n} a_i z^i$ un polinomio de grado $n \geq 1$ con coeficientes complejos, entonces $\exists r \in \mathbb{C}$ de modo que $P_n(r) = 0$

La demostración de este teorema se omite pues excede, en mucho, el nivel de este curso. ▼

Teorema 5.5.2 (Teorema del resto)
Sea $P_n(x)$ un polinomio de grado $n \geq 1$ en la indeterminada x, y sea $a \in \mathbb{C}$. Entonces $P_n(a)$ coincide con el resto de dividir $P_n(x)$ por $x - a$.

Dem:

$P_n(x) = Q_{n-1}(x)(x-a) + r$, donde $Q_{n-1}(x)$ es el cociente y $r \in \mathbb{C}$ el resto. Entonces se tiene que $P_n(a) = Q_{n-1}(a)(a-a) + r = r$. ▼

Teorema 5.5.3 (Teorema de Rufini)
Sea $P_n(x)$ un polinomio de grado $n \geq 1$ en la indeterminada x, y sea $a \in \mathbb{C}$. Entonces $P_n(x)$ es divisible por $x - a$ si, y sólo si, $P_n(a) = 0$.

Dem:

Consecuencia inmediata del anterior, pues $P_n(x)$ divisible por $x - a$ si, y sólo si, el resto $r = 0$, pero $r = P_n(a)$ de donde se sigue el resultado. ▼

Teorema 5.5.4
Sea $P_n(z) = \sum_{i=0}^{n} a_i z^i$ un polinomio de grado $n \geq 1$ con coeficientes complejos, entonces $\exists r_1, \ldots, r_s \in \mathbb{C}$, y $\exists n_1, \ldots, n_s \in \mathbb{N}$ de modo que $P_n(z) = a_n \prod_{i=1}^{s} (z - r_i)^{n_i}$ con $\sum_{i=1}^{s} n_i = n$.

Dem:

Consecuencia inmediata del teorema fundamental del álgebra y del teorema de Rufini. ▼

Teorema 5.5.5
Sea $P_n(z) = \sum_{i=0}^{n} a_i z^i$ un polinomio de grado n con coeficientes en $\mathbb{C}$ y sea $r \in \mathbb{C} - \{0\}$ una raíz de $P_n(z)$. Entonces $\frac{1}{r}$ es raíz del polinomio $Q_n(z) = \sum_{i=0}^{n} a_i z^{i-n}$.

Dem:

$P_n(z) = \sum_{i=0}^{n} a_i z^i$, entonces $\forall z \neq 0$ se tiene

$$P_n(z) = z^n \sum_{i=0}^{n} a_i z^{i-n} = z^n Q_n(z).$$

Por tanto, si $r \neq 0$ es raíz de $P_n(z)$ se tiene que $0 = P_n(r) = r^n Q_n(\frac{1}{r})$ y puesto que $r \neq 0$ $Q_n(\frac{1}{r}) = 0$, con lo que se tiene el resultado. ▼

Teorema 5.5.6
Sea $P_n(x)$ un polinomio de grado $n \geq 1$ con coeficientes reales, y sea $r \in \mathbb{C}$ raíz de $P_n(x)$, entonces $\overline{z}$ también es raíz de $P_n(x)$.

Dem:

Sea $r \in \mathbb{C}$ raíz de $P_n(x)$, entonces $P_n(r) = \sum_{i=0}^{n} a_n r^n = 0$. Por otra parte, un número complejo z es real si, y sólo si, $z = \overline{z}$ donde $\overline{z}$ es la conjugación compleja, ésto es $\overline{\alpha + i\beta} = \alpha - i\beta$.

Tomando conjugados en ambos miembros de la igualdad $P_n(r) = 0$ se tiene que $\overline{P_n(r)} = \overline{0} = 0$, y por tanto

$$\overline{P_n(r)} = \overline{\sum_{i=0}^{n} a_n x^n} = \sum_{i=0}^{n} \overline{a_n}\, \overline{r}^n = \sum_{i=0}^{n} a_n \overline{r}^n = P_n(\overline{r})$$

y por tanto, $P_n(\overline{r}) = 0$, de donde se sigue el resultado. ▼

Teorema 5.5.7
Sea $P_n(x)$ un polinomio de grado $n \geq 1$ con coeficientes reales. Si $z = \alpha + i\,\beta$ es raíz, entonces $P_n(x)$ es divisible por $(x - \alpha)^2 + \beta^2$.

Dem:

Por el teorema anterior si $\alpha + i\,\beta$ es raíz de $P_n(x)$, entonces $\alpha - i\,\beta$ también lo es. Por tanto, $P_n(x)$ es divisible por $x - [\alpha + i\,\beta]$ y por $x - [\alpha - i\,\beta]$, por lo que es divisible por el producto de ambos. Así, $P_n(x)$ es divisible por

$$(x - [\alpha + i\,\beta])\,(x - [\alpha - i\,\beta]) = (x - \alpha)^2 - i^2\beta^2 = (x - \alpha)^2 + \beta^2.$$

de donde se sigue el teorema. ▼

Teorema 5.5.8
Sea $P_n(x)$ un polinomio de grado $n \geq 1$ con coeficientes reales. Entonces $P_n(x)$ se factoriza como

$$P_n(x) = K \prod_{i=1}^{s} (x - x_i)^{n_i} \prod_{j=1}^{r} \left[(x - \alpha_j)^2 + \beta_j \right]^{m_j}$$

donde $K \in \mathbb{R}$, $x_1, \ldots, x_s \in \mathbb{R}, \alpha_1, \ldots, \alpha_r, \beta_1, \ldots, \beta_r \in \mathbb{R}, n_1, \ldots, n_s, m_1, \ldots, m_r \in \mathbb{N}$, $n_1 + \ldots + n_s + 2m_1 + \ldots + 2m_r = n$.

Dem:

Se deduce inmediatamente de los resultados anteriores. ▼

Definición 5.5.2
Sea $P_n(x)$ un polinomio cuya factorización completa viene dada por

$$P_n(x) = K \prod_{i=1}^{s} (x - x_i)^{n_i} \prod_{j=1}^{r} \left[(x - \alpha_j)^2 + \beta_j \right]^{m_j}.$$

Entonces a los números $n_1, \ldots, n_s, m_1, \ldots, m_r$ se les denomina **grado de multiplicidad de las raíces** $r_1, \ldots, r_s$ y $\alpha_1 \pm i\,\beta_1, \ldots, \alpha_r \pm i\,\beta_r$, respectivamente.

Teorema 5.5.9
Sea $P_n(z)$ un polinomio con coeficientes complejos y sea $r \in \mathbb{C}$ raíz de multiplicidad $1 \leq k \leq n$. Entonces r es raíz de $P_n^{k)}(z)$, $i = 0, \ldots, k-1$ y no lo es de $P_n^{k)}(z)$ donde $P_n^{i)}(z) = \frac{d^i}{duz^i} P_n(z)$.

Dem:

Si r es raíz de orden k de $P_n(z)$, entonces $P_n(z) = (z-r)^k Q(z)$ donde $Q(z)$ es un polinomio de grado $n-k$ de manera que $Q(r) \neq 0$. Si $i \leq k$ se tiene que

$$P_n^{i)}(z) = \sum_{j=0}^{i} \binom{i}{j} \left(\frac{d^j}{dz^j}(z-r)^k \right) \left(\frac{d^{i-j}}{dz^{i-j}} Q(z) \right),$$

así

$$P_n^{i)}(z) = \sum_{j=1}^{i} \binom{i}{j} \frac{k!}{(k-j)!}(z-r)^{k-j} Q^{i-j}(z.)$$

Por tanto, si $i < k$ se tiene que $P_n^{i)}(z) = (z-r)^{k-i} R(z)$ donde $R(z)$ es un polinomio en z, de donde $P_n(r) = 0$. Por otra parte, $P_n^{k)}(z) = (z-r)S(z) + k!Q(z)$ siendo $S(z)$ un polinomio en z, de donde $P_n^{k)}(r) = (r-r)S(r) + k!Q(r) = k!Q(r) \neq 0$ pues $Q(r) \neq 0$.

▼

Teorema 5.5.10

Sea $P_n(z)$ un polinomio con coeficientes complejos y sea $r \in$ raíz de multiplicidad $k > 1$. Entonces r es también raíz de $P_n'(z)$ de multiplicidad $k-1$.

Dem:

$P_n(z) = (z-r)^k Q(z)$ donde $Q(z)$ es un polinomio en z de modo que $Q(r) \neq 0$. Entonces derivando se tiene

$$P_n'(z) = k(z-r)^{k-1} Q(z) + (z-r)^k Q'(z)$$

y por tanto:

$$P_n'(z) = (z-r)^k \left[kQ(z) + (z-r)Q'(z) \right] = (z-r)^{k-1} S(z)$$

donde $S(z) = kQ(z) + (z-r)Q'(z)$.

Por otra parte, $S(r) = kQ(r) + (r-r)Q'(r) = kQ(r) \neq 0$, de donde se sigue el resultado.

▼

Definición 5.5.3

Sean $P(z)$, $Q(z)$ polinomios de coeficientes en $\mathbb{C}$. Diremos que un polinomio $D(z)$ con coeficientes en $\mathbb{C}$ es **máximo común divisor** de $P(z)$ y $Q(z)$ si $D(z)$ divide a $P(z)$ y a $Q(z)$ y para cualquier otro divisor común $R(z)$ de $P(z)$ y $Q(z)$ se verifica que $\mathrm{grad}(R(z)) \leq \mathrm{grad}(D(z))$.

Diremos que los polinomios $P(z)$ y $Q(z)$ son primos entre sí cuando $\mathrm{grad}(D(z)) = 0$.

Teorema 5.5.11

Sea el polinomio $D(z)$ divisor común de los polinomios $P(z)$ y $Q(z)$ con $\mathrm{grad}(P(z) \leq \mathrm{grad}(Q(z))$, y sea $r(z)$ el resto de dividir $P(z)$ entre $Q(z)$. Entonces $D(z)$ divide a $r(z)$.

Dem:

Sea $c(z)$ el cociente de dividir $P(z)$ entre $Q(z)$, entonces $P(z) = Q(z)c(z) + r(z)$. Puesto que $D(z)$ divide a $P(z)$ se tiene que $P(z) = D(z)p(z)$; por la misma razón se tiene $Q(z) = D(z)q(z)$. Por tanto, $D(z)p(z) = D(z)q(z)c(z) + r(z)$, de donde $r(z) = D(z)\left[p(z) - q(z)c(z)\right]$. Por tanto, $D(z)$ divide a $r(z)$. ▼

El cálculo del máximo común divisor de dos polinomios puede efectuarse fácilmente por medio del algoritmo de Euclides. Este algoritmo está basado en el teorema anterior y consiste en lo siguiente:

1. Tomar el polinomio de mayor grado como $D(z)$ y el de menor como $d(z)$. En caso de ser del mismo grado tomar como $D(z)$ y el otro como $d(z)$.

2. Efectuar la división de $D(z)$ entre $d(z)$. Sea $c(z)$ el cociente y $r(z)$ el resto de modo que $D(z) = c(z)d(z) + r(z)$.

3. Si $r(x) = 0$, el máximo común divisor buscado es $d(z)$, con lo que finaliza el algoritmo. Si por el contrario $r(z) \neq 0$ hacer $D(z) = d(z)$ y $d(z) = r(z)$ e ir al paso 2.

Ejemplo 5.5.1

Calcular el máximo común divisor de los polinomios $z^4 + 3z^3 + 4z^2 + 3z + 1$ y $z^3 + 2z^2 - z - 2$.

Tomando como $D(z) = z^4+3z^3+4z^2+3z+1$ y como $d(z) = z^3+2z^2-z-2$ y efectuando la división, se tiene

$$D(z) = d(z)c(z) + r(z), \quad \text{donde } c(x) = z + 1,\ r(x) = 3z^2 + 6z + 3.$$

Puesto que $r(x) \neq 0$, sea $D(z) = z^3 + 2z^2 - z - 2$ y $d(z) = 3z^2 + 6z + 3$, entonces se tiene

$$D(z) = d(z)c(z) + r(z), \quad \text{donde } c(x) = \frac{z}{3},\ r(x) = -2z - 2.$$

Puesto que $r(x) \neq 0$ sea $D(z) = 3z^2 + 6z + 3$ y $d(z) = -2z - 2$, entonces se tiene

$$D(z) = d(z)c(z) + r(z), \quad \text{donde } c(x) = -\frac{3}{2}z - \frac{3}{2},\ r(x) = 0.$$

Puesto que $r(x) = 0$, el máximo común divisor es $d(z) = -2z - 2$.

También es posible dar como máximo común divisor $k\,d(z)$ donde $k \in \mathbb{R} - \{0\}$; así, podemos dar como resultado $z + 1$.

El algoritmo de Euclides puede aplicarse de forma tabular como:

	$z+1$	$\frac{z}{3}$	$-\frac{3}{2}z - \frac{3}{2}$
$z^4 + 3z^3 + 4z^2 + 3z + 1$	$z^3 + 2z^2 - z - 2$	$3z^2 + 6z + 3$	$-2z - 2$
	$3z^2 + 6z + 3$	$-2z - 2$	0

♦

Teorema 5.5.12
Sea $P(z)$ un polinomio de grado $n \geq 1$ y sea $D(z)$ el máximo común divisor de $P_n(z)$ y de su derivada $P_n'(z)$. Entonces, el polinomio $\frac{P_n(z)}{D(z)}$ tiene las mismas raíces que $P_n(z)$ pero con multiplicidad uno.

Dem:
Sea un polinomio de grado $n \geq 1$. Entonces $P_n(z)$ se factoriza como

$$P_n(x) = K \prod_{i=1}^{s} (z - x_i)^{n_i} \prod_{j=1}^{r} \left[(z - \alpha_j)^2 + \beta_j \right]^{m_j} ,$$

por tanto

$$P_n'(z) = S(z) \prod_{i=1}^{s} (z - x_i)^{n_i - 1} \prod_{j=1}^{r} \left[(z - \alpha_j)^2 + \beta_j \right]^{m_j - 1} ,$$

donde $S(z)$ es un polinomio que no tiene por raíces x_i, $i = 1, \ldots, s$, $\alpha_i \pm i\,\beta_i$, $j = 1, \ldots, r$, por tanto, el máximo común divisor de $P_n(z)$ y $P_n'(z)$ es

$$D(z) = \prod_{i=1}^{s} (z - x_i)^{n_i - 1} \prod_{j=1}^{r} \left[(z - \alpha_j)^2 + \beta_j \right]^{m_j - 1} ,$$

y por tanto

$$\frac{P_n(z)}{D(z)} = \prod_{i=1}^{s} (z - x_i) \prod_{j=1}^{r} \left[(z - \alpha_j)^2 + \beta_j \right] ,$$

con lo que se tiene el resultado.

5.5.1. Acotación y separación de raíces

En el caso de ecuaciones polinómicas es posible establecer, por una parte, cotas superiores e inferiores a los valores de sus raíces, y también en el caso de ser éstas reales simples, determinar intervalos en los que se encuentra una única raíz, técnica conocida como separación de raíces.

Teorema 5.5.13
Sea la ecuación polinómica $P_n(z) = \sum_{i=0}^{n} a_i x^i = 0$ y sea $A = \max\{|a_0|, \ldots, |a_{n-1}|\}$. Entonces si $r \in \mathbb{C}$ es raíz de $P_n(z)$, se verifica que $|r| < 1 + \frac{A}{a_n}$.

Dem:
Si $|z| > 1$ se tiene:

$$|P_n(z)| = |a_n z^n + a_{n-1} z^{n-1} + \ldots + a_1 z + a_0| \geq$$

$$\geq |a_n z^n| - |a_{n-1} z^{n-1} + \ldots + a_1 z + a_0| \geq$$

$$\geq |a_n z^n| - A \left[|z|^{n-1} + \ldots + |z| + 1 \right] =$$

$$= |a_n||z|^n - A\frac{|z|^n - 1}{|z| - 1} > \left[|a_n| - \frac{A}{|z| - 1}\right]|z|^n.$$

Por tanto, si

$$|a_n| - \frac{A}{|z| - 1} \geq 0$$

z no puede ser raíz de la ecuación, de donde para que r con $|r| > 1$ sea raíz es necesario que

$$|a_n| - \frac{A}{|r| - 1} < 0,$$

lo que es equivalente a

$$|r| < 1 + \frac{A}{a_n}.$$

Si $|r| < 1$, la relación se verifica directamente. Finalmente, si $|r| = 1$, también se verifica pues $A > 0$ ya que en caso de ser $A = 0$ la única raíz de la ecuación sería $r = 0$. ▼

Teorema 5.5.14
Sea la ecuación polinómica dada por $P_n(z) = \sum_{i=0}^{n} a_i x^i = 0$ con $a_0 \neq 0$ y sea $B = \text{máx}\{|a_1|, \ldots, |a_n|\}$. Entonces si $r \in \mathbb{C}$ es raíz de la ecuación $P_n(z) = 0$ se verifica $|r| > \frac{1}{1+\frac{B}{|a_0|}}$.

Dem:

Basta considerar que el polinomio $Q_n(z) = z^n P_n(\frac{1}{z})$ del cual es raíz $\frac{1}{r}$ y cuyo desarrollo viene dado por $Q_n(z) = a_0 z^n + \ldots + a_{n-1}z + a_n$, por tanto $\frac{1}{r} < 1 + \frac{B}{a_0}$ donde $B = \text{máx}\{|a_1|, \ldots, |a_n|\}$, a partir de lo cual se tiene

$$|r| > \frac{1}{1 + \frac{B}{|a_0|}}.$$

▼

Ejemplo 5.5.2
Sea el polinomio $P_n(z) = z^5 + 3z^4 - 7z^3 + z^2 - 1$. Encontrar una acotación para las raíces complejas de la ecuación $P_n(z) = 0$.

Sean $A = \text{máx}\{|a_0|, \ldots, |a_{n-1}|\}$, $B = \text{máx}\{|a_1|, \ldots, |a_n|\}$. Entonces se verifica que si $r \in \mathbb{C}$ es raíz de la ecuación $P_n(z) = 0$, debe verificarse que

$$\frac{1}{1 + \frac{B}{|a_0|}} < |r| < 1 + \frac{A}{|a_n|},$$

y puesto que en este caso $A = B = 7$ se tiene que

$$\frac{1}{1 + \frac{7}{1}} < |r| < 1 + \frac{7}{1},$$

de donde se obtiene

$$\frac{1}{8} < |r| < 7.$$

♦

En el caso de polinomios con coeficientes reales es posible proporcionar límites mas estrechos para las raíces, así como establecer cual es el número de raíces reales comprendidas en un intervalo $[a, b]$. Para ello, en primer lugar, sea $P_n(x) = a_n x^n + \cdots + a_1 x + a_0$ un polinomio de grado $n \geq 1$ con coeficientes reales, y considérese los polinomios auxiliares $S_n(x) = P_n(-x)$, y también si $a_0 \neq 0$ $Q_n(x) = x^n P_n(\frac{1}{x})$. El primero de dichos polinomios tiene por raíces las mismas de $P_n(x)$ cambiadas de signo, siendo las del segundo las inversas de las raíces de $P_n(x)$.

Teorema 5.5.15 (Teorema de Lagrange)
Sea $P_n(x) = a_n x^n + \ldots + a_1 x + a_0$ un polinomio de grado $n \geq 1$ y sea x^+ una raíz positiva de $P_n(x) = 0$. Entonces si $|a_n| > 0$, $a_k < 0$ es el coeficiente negativo de mayor grado y $B = \max\{|a_i| : a_i < 0,\ i = 0, \ldots, n-1\}$, se verifica

$$x^+ < 1 + \sqrt[n-k]{\frac{B}{a_n}}$$

.

Dem:

Si $x^+ \leq 1$ se tiene el teorema (de modo totalmente similar a como se obtuvo en el teorema anterior).

Si $x^+ > 1$, sea $I = \{i \in \{1, 2, \ldots, n\} : a_i > 0\}$ y sea $\overline{I} = \{1, \ldots, n\} - I$. Entonces $\forall x > 0$:

$$P_n(x) = \sum_{i \in I} a_i x^i + \sum_{i \in \overline{I}} a_i x^i \geq a_n x^n - B \sum_{i \in \overline{I}} x^i \geq$$

$$\geq a_n x^n - B \sum_{i=0}^{n-k} x^i = a_n x^n - B \frac{x^{n-k+1} - 1}{x - 1}.$$

Puesto que $x > 1$ se tiene que

$$P_n(x) = \frac{x^{k+1}}{x-1} \left[a_n x^{n-k-1}(x-1) - B \right] > \frac{x^{k+1}}{x-1} \left[a_n (x-1)^{n-k} - B \right],$$

y por tanto, si $x > 0$ es raíz de $P_n(x)$, no puede darse la condición

$$a_n (x-1)^k - B \geq 0$$

de donde $a_n (x^+)^{n-k} - B < 0$, y por tanto:

$$x^+ < 1 + \sqrt[n-k]{\frac{B}{a_n}}.$$

▼

Ejemplo 5.5.3
Sea el polinomio $P_n(z) = z^5 + 3z^4 - 7z^3 + z^2 - 1$. Encontrar una acotación para las raíces reales de la ecuación $P_n(z) = 0$.

Sea x^+ raíz real positiva de $P_n(x) = 0$. El coeficiente negativo correspondiente al mayor grado es $a_3 = -7$ y $B = \text{máx}\{|a_i| : a_i < 0, i = 0, \ldots, n-1\} = 7$. Entonces, se tiene

$$x^+ < 1 + \sqrt[5-3]{\frac{B}{a_n}} = 1 + \sqrt{7} < 1 + 2.65 = 3.65.$$

Para establecer una cota inferior de las raíces positivas sea el polinomio $\pm Q_n(x) = x^n P_n(\frac{1}{x}) = -x^5 + x^3 - 7x^2 + 3x + 1$, que multiplicado por -1 para que el coeficiente principal sea positivo, resulta $Q_n(x) = x^5 - x^3 + 7x^2 - 3x - 1$; así, en este caso $k = 3$ y $B = 3$. Por tanto:

$$\frac{1}{x^+} < 1 + \sqrt[5-3]{\frac{3}{1}} = 1 + \sqrt{3} < 2.74.$$

Por tanto, se tiene

$$x^+ > \frac{1}{2.74} > 0.3.$$

Finalmente, podemos asegurar que las raíces positivas satisfacen $0.36 < x^+ < 3.6.5$.

Respecto a las raíces reales negativas considérese el polinomio $\pm S_n(z) = P_n(-x)$. Tomando el signo $-$ para que el coeficiente principal sea positivo, resulta $S_n(z) = x^5 - 3x^4 + 7x^3 - x^2 + 1$, y aplicando el teorema anterior se tiene que

$$-x^- < 1 + \sqrt[5-4]{\frac{3}{1}} = 1 + 3 = 4,$$

por tanto $x^- > -4$. Finalmente tomando el polinomio $x^5 - x^3 + 7x^2 - 3x + 1$ se tiene que

$$-\frac{1}{x^-} < 1 + \sqrt[5-3]{\frac{3}{1}} = 1 + \sqrt{3} < 2.74,$$

de donde $x^- < -\frac{1}{2.74} < -0.36$. Por tanto:

$$-4 < x^- < -0.36.$$

Este procedimiento aporta habitualmente mejores acotaciones que el método general. ♦

Teorema 5.5.16 (Teorema de Newton)
Sea $P_n(x) = \sum_{i=0}^{n} a_i x^i$ un polinomio de grado $n \geq 1$ de coeficientes reales con $a_n > 0$ y sea $c \in \mathbb{R}^+$ de modo que $P_n(c) > 0$, $P_n^{k)}(c) > 0$, $k = 1, \ldots, n$. Entonces si $x^+ \in \mathbb{R}^+$ es raíz de la ecuación $P_n(x) = 0$, se verifica que $x^+ < c$.

Dem:
Sea

$$P_n(x) = P_n(c) + \sum_{k=1}^{n} \frac{P_n^{k)}}{k!}(c)(x-c)^k,$$

el desarrollo en serie de Taylor del polinomio $P(x)$ alrededor del punto c. Entonces si $x \geq c$ se tiene que

$$P_n(x) = P_n(c) + \sum_{k=1}^{n} \frac{P_n^{k)}}{k!}(c)(x-c)^k \geq P_n(c) + \sum_{k=1}^{n} \frac{P_n^{k)}}{k!}(c-c)^k = P_n(c) > 0$$

Por tanto, $x \geq c$ no puede ser raíz de $P_n(x)$. Entonces si $x^+ \in \mathbb{R}^+$ satisface $P_n(x^+) = 0$, se tiene que $x^+ < c$. ▼

Ejemplo 5.5.4
Sea el polinomio $P_n(z) = x^5 + 3x^4 - 7x^3 + x^2 - 1$. Encontrar una acotación para las raíces reales positivas de la ecuación $P_n(z) = 0$.

En primer lugar se calculan las derivadas de $P_n(x)$; así, se tiene

$$\begin{aligned}
P_n^{1)}(x) &= 5x^4 + 12x^3 - 21x^2 + 2x \\
P_n^{2)}(x) &= 20x^3 + 36x^2 - 42x + 2 \\
P_n^{3)}(x) &= 60x^2 - 72x - 42 \\
P_n^{4)}(x) &= 120x - 72 \\
P_n^{5)}(x) &= 120.
\end{aligned}$$

En la siguiente tabla se da el valor de $P_n(c)$ y de sus derivadas para distintos valores de c, comenzando por el valor 2.74 proporcionado por la acotación de Lagrange.

c	$P_n(c)$	$P_n'(c)$	$P_n''(c)$	$P_n^{3)}(c)$	$P_n^{4)}(c)$	$P_n^{5)}(c)$
2.74	186.04	376.49	568.61	605.74	400.80	120.00
2.70	171.43	354.23	544.70	589.80	396.00	120.00
2.50	110.72	256.56	434.50	513.00	372.00	120.00
2.30	67.44	179.43	339.18	441.00	348.00	120.00
2.10	37.77	119.96	257.78	373.80	324.00	120.00
1.90	18.45	75.46	189.34	311.40	300.00	120.00
1.70	6.75	43.43	132.90	253.80	276.00	120.00
1.50	0.41	21.56	87.50	201.00	252.00	120.00
1.30	-2.41	7.75	52.18	153.00	228.00	120.00

Por tanto, podemos concluir que $x^+ < 1.5$. ♦

5.5.2. Número de raíces. Separación

Dado un polinomio $P_n(x)$ de grado $n \geq 1$ resulta de interés conocer, por una parte, el número de raíces reales de la ecuación en un cierto intervalo, y por otra, ser capaces de encontrar intervalos en los que exista una única raíz. Este último proceso se conoce como separación de raíces.

En primer lugar, recordar que dado un polinomio $P_n(x)$ se tiene que

- Si dado un intervalo $[a, b]$, $a < b$ se verifica $P_n(a)P_n(b) < 0$. Entonces $P_n(x)$ tiene un número impar de raíces en $[a, b]$.

- Si dado un intervalo $[a, b]$, $a < b$ se verifica $P_n(a)P_n(b) > 0$. Entonces $P_n(x)$ o no tiene raíces en $[a, b]$ o tiene un número par de raíces en $[a, b]$.

Definición 5.5.4
Sea una n-tupla de números distintos de cero $(c_1, c_2, \ldots, c_n)$. Llamamos **número de cambios de signo** de $(c_1, c_2, \ldots, c_n)$ al cardinal de conjunto $N = \{i \in \{1, 2, \ldots, n-1\} : c_i c_{i+1} < 0\}$. Si existen elementos nulos, también es posible definir número de cambios de signo como el número de cambios de signo de la r-tupla $(c_{i_1}, \ldots, c_{i_r})$ formada por los elementos no nulos, donde el orden se conserva.

Definición 5.5.5
Sean una n-tupla de números $(c_1, c_2, \ldots, c_n)$ con $c_1 \neq 0$, $c_n \neq 0$. Llamamos **número superior de cambios de signo** de $(c_1, c_2, \ldots, c_n)$ al cardinal del conjunto $\overline{N} = \{i \in \{1, 2, \ldots, n-1\} : a_i a_{i+1} < 0\}$ donde $a_i = c_i$ si $c_i \not 0$ y si $c_k = c_{k+1} = \ldots, c_{k+1} = 0$ con $c_{k-1} \neq 0$, $c_{k+l} \neq 0$ $a_{k+i} = (-1)^{l-i} c_{k+l}$, $i = 0, \ldots, l-1$.

Definición 5.5.6
Sea una n-tupla de números $(c_1, c_2, \ldots, c_n)$ con $c_1 \neq 0$, $c_n \neq 0$. Llamamos **número inferior de cambios de signo** de $(c_1, c_2, \ldots, c_n)$ al cardinal de conjunto $\underline{N} = \{i \in \{1, 2, \ldots, n-1\} : a_i a_{i+1} < 0\}$ donde $a_i = c_i$ si $c_i \neq 0$ y si $c_k = c_{k+1} = \ldots, c_{k+l-1} = 0$, $a_{k+i} = c_{k-1}$, $i = 0, \ldots, l-1$.

Ejemplo 5.5.5
Dada la n-tupla $(2, 0, 0, -1, 0, -2, 2)$, se tiene que

- $\overline{N}$ es el número de cambios de signo de $(2, -2, 2, -1, 1, -2, 2)$ así $\overline{N} = 6$.
- $\underline{N}$ es el número de cambios de signo de $(2, 2, 2, -1, -1, -2, 2)$ así $\underline{N} = 2$.

♦

Definición 5.5.7
Sea la secuencia formada por los polinomios $f_0(x), f_1(x), \ldots, f_n(x)$. Diremos que dicha secuencia es una **secuencia de Sturm** si:

1. Los polinomios consecutivos de la secuencia anterior no tienen raíces comunes.
2. $f_n(x)$ es una constante.
3. Si α_k es raíz de $f_k(x)$ con $k = 1, \ldots, n-1$ se tiene que $f_{k-1}(\alpha_k) f_{k+1}(\alpha_k) < 0$.
4. Si α_0 es raíz de $f_0(x)$, se tiene que el producto $f_0(x) f_1(x)$ cambia de signo de $-$ a $+$ cuando al crecer x pasa por $x = \alpha_0$.

Definición 5.5.8
Sea $P_n(x)$ un polinomio de grado $n \geq 1$. Llamamos secuencia de Sturm del polinomio $P_n(x)$ a la secuencia $(S_0(x), S_1(x), S_2(x), \ldots, S_n(x))$ definida como $S_0(x) = P_n(x)$, $S_1(x) = P'_n(x)$ y para $k \geq 2$, $S_k(x)$ es el resto cambiado de signo resultante dividir $S_{k-2}(x)$ entre $S_{k-1}(x)$.

Teorema 5.5.17
Sea $P_n(x)$ un polinomio de grado $n \geq 1$ el cual no contiene raíces múltiples. Entonces la secuencia anterior es efectivamente una secuencia de Sturm.

Dem:

1. En efecto, si $x = \alpha$ raíz de $S_k(x)$ y $S_{k+1}(x)$ con $k = 0$ se tiene que α es raíz común de $P_n(x)$ y de $P_n'(x)$, por tanto, es raíz de multiplicidad mayor o igual que 2 de $P_n(x)$, lo cual es absurdo. Si $k > 0$, sea k el menor valor para el que se tiene que α es raíz común de $S_k(x)$ y $S_{k+1}(x)$, entonces
$$S_{k-1}(x) = S_k(x)Q_k(x) - S_{k+1}(x)$$
por tanto, α también es raíz de $S_{k-1}(x)$, lo cual es imposible por la condición de mínimo de k.

2. Puesto que $P_n(x)$ no contiene raíces múltiples, $P_n(x)$ y $P_n'(x)$ son primos entre sí lo que implica que su máximo común divisor es una constante. Según la construcción de la secuencia este máximo común divisor es $S_n(x)$ pues el tomar signo menos para los restos no modifica el algoritmo de Euclides.

3. $S_{k-1}(x) = S_k(x)Q_(x) - S_{k+1}(x)$ por tanto si $S_k(\alpha) = 0$ se tiene que $S_{k-1}(\alpha) = -S_{k-1}(\alpha)$ y puesto que $S_k(\alpha \neq 0$ se tiene $S_{k-1}(\alpha)S_{k+1}(\alpha) < 0$.

4. Inmediato, pues si $S_0(\alpha) = 0$ $\exists \epsilon > 0$ de modo que $S_0(x) \neq 0$ si $x \in [\alpha - \epsilon, \alpha + \epsilon] - \{\alpha\}$ y $P_n'(x) \neq 0$ tiene signo constante si $x \in [\alpha - \epsilon, \alpha + \epsilon]$. Entonces si $x \in [\alpha - \epsilon, \alpha[$ y $S_0(x) > 0$ se tiene que $P_n(x) > 0$ por tanto, debe ser decreciente de donde $P_n'(x) < 0$, así $S_1(x) < 0$, y por tanto $S_0(x)S_1(x) < 0$ si $x \in [\alpha - \epsilon, \alpha[$, por otra parte, en $]\alpha, \alpha + \epsilon]$ $P_n(x) < 0$ y, por tanto, $S_0(x)S_1(x) > 0$. Si $P_n(x) < 0$ si $x \in [\alpha - \epsilon, \alpha[$, $P_n'(x) > 0$ pues $P_n(x)$ debe ser creciente en $[\alpha - \epsilon, \alpha + \epsilon]$ con lo que también se tiene que $S_0(x)S_1(x) < 0$ si $x \in [\alpha - \epsilon, \alpha[$ y $S_0(x)S_1(x) > 0$ si $x \in]\alpha, \alpha + \epsilon]$.

▼

Ejemplo 5.5.6
Sea el polinomio $P(x) = x^5 - 2x^3 + x - 1$. Hallar su secuencia de Sturm.

Para hallar la secuencia de Sturm, sea $S_0(x) = x^5 - 2x^3 + x - 1$. $S_1(x) = P_n'(x) = 5x^4 - 6x^2 + 1$. Así, efectuando la división se tiene que

$$S_0(x) = S_1(x)\left(\frac{5}{4}x^2 - \frac{5}{4}\right) + \left(-\frac{5}{4}x^2 + \frac{1}{4}\right) \qquad S_2(x) = \frac{5}{4}x^2 - \frac{1}{4}$$
$$S_1(x) = S_2(x)\left(\frac{16}{25}x\right) + \left(-\frac{16}{25}x + 1\right) \qquad S_3(x) = \frac{16}{25}x - 1$$
$$S_2(x) = S_3(x)\left(\frac{125}{64}x + \frac{3125}{1024}\right) + \frac{2869}{1024} \qquad S_4(x) = -\frac{2869}{1024}$$

Por tanto, la secuencia de Sturm es

$$\left\{x^5 - 2x^3 + x - 1, 5x^4 - 6x^2 + 1, \frac{5}{4}x^2 - \frac{1}{4}, \frac{16}{25}x - 1, -\frac{2869}{1024}\right\}$$

♦

Teorema 5.5.18 (Teorema de Sturm)
Sea $P_n(x)$ un polinomio de grado $n \geq 1$ el cual no contiene raíces múltiples. Sea $N(x)$ el número de variaciones de signo en la secuencia de Sturm $S_0(x), \ldots, S_n(x)$ del polinomio, sea $[a, b] \subset \mathbb{R}$, $a < b$ y sea $N(a, b)$ el número de raíces reales de $P_n(x)$ contenidas en dicho intervalo $[a, b]$. Entonces $N(a, b) = N(a) - N(b)$.

Dem:
Sea $N(c)$ en número de variaciones de signo en la secuencia de Sturm del polinomio $P_n(x)$ para $x = c$. El valor $N(c)$ sólo puede variar al pasar por una raíz de uno de los polinomios $S_k(x)$ de la secuencia de Sturm. Pueden darse dos casos: $k = 0$ y $k \neq 0$.

Si $k = 0$, $\exists \epsilon > 0$, de modo que si $x_1 \in [c - \epsilon, c[$, $S_0(x_1)S_1(x_1) < 0$ y $S_0(x_2)S_1(x_2) > 0$, cuando $x_2 \in]c, c + \epsilon]$. Por tanto, en los tres primeros términos de la secuencia de Sturm hay un cambio más de signo en x_1 que en x_2, pues únicamente pueden darse para $S_0(c - \epsilon), S_1(c - \epsilon); S_0(c + \epsilon), S_1(c + \epsilon)$ los casos $+, -, ; -, -,$ y $-, +; +, +$.

Si $k \neq 0$ y c es raíz de $S_k(x)$, por ser $S_0(x), \ldots, S_n(x)$ secuencia de Sturm se tiene que $S_{k-1}(c)S_{k-1}(c) < 1$. Por otra parte, $\exists \epsilon > 0$ de modo que $S_{k-1}(x) \neq 0$ y $S_{k+1}(x) \neq 0$ si $x \in [c - \epsilon, c + \epsilon]$ y c es la única raíz de $S_k(x)$ en $[c - \epsilon, c + \epsilon]$. De este modo, la secuencia formada por $S_{k-1}(x), S_k(x), S_{k+1}(x)$ no presenta cambios de signo en $[c - \epsilon, c + \epsilon]$ pues para las ternas

$$S_{k-1}([c - \epsilon), S_k([c - \epsilon), S_{k+1}([c - \epsilon); S_{k-1}([c + \epsilon), S_k([c + \epsilon), S_{k+1}([c + \epsilon)$$

sólo caben las posibilidades

$$+, +, -; +, -, - \quad +, -, -; +, +, - \quad -, +, +; -, -, + \quad -, -, +; -, +, +$$

▼

Dado que el método de Sturm puede resultar complicado enunciaremos a continuación algunos resultados acerca del número de raíces de un polinomio, resultados que en ocasiones pueden sernos útiles.

Teorema 5.5.19 (Teorema de Boundan-Fourier)
Sea $P_n(x)$ un polinomio de grado $n \geq 1$ con coeficientes reales, y sean $a, b \in \mathbb{R}$, $a < b$ de modo que $P_n(a)P_n(b) \neq 0$, y sea $N(a, b)$ el número de raíces de $P_n(x)$ en el intervalo $[a, b]$ donde cada raíz cuenta tantas veces como su multiplicidad. Sea la secuencia $(P_n(x), P'_n(x), \ldots, P_n^{n)}(x))$, y sea $\overline{N}(x)$, $\underline{N}(x)$ los números superior e inferior de cambios de signo de la secuencia anterior. Entonces se verifica que

$$N(a, b) = \underline{N}(a) - \overline{N}(b) - 2k, k = 0, 1, \ldots E\left[\frac{\underline{N}(a) - \overline{N}(b)}{2}\right]$$

donde $E[]$ representa la función parte entera.

Dem:
Considérese la secuencia $P_n(x), P'_n(x), \ldots, P_n^{n)}(x)$. Sean $a, b \in \mathbb{R}$ con $a < b$ de modo que $P_n(a) \neq 0$, $P_n(b) \neq 0$. Sea $\alpha \in [a, b]$. Si α no es raíz de ninguno de los polinomios de la secuencia, se tiene que $\exists \epsilon > 0$ de modo que ninguno de

los polinomios de la secuencia se anula en el intervalo $[\alpha-\epsilon,\alpha+\epsilon]$ y, por tanto, $N(\alpha-\epsilon)=N(\alpha+\epsilon)$.

Si α es raíz de orden k de $P_n(x)$ con $1\leq k\leq n$ se tiene que $P_n(\alpha)=P'_n(\alpha)=\cdots=P_n^{k-1)}(\alpha)=0$ y $P_n^{k)}(\alpha)\neq 0$. Sea $\epsilon>0$ de modo que $P_n^{s)}(x)\neq 0$ si $x\in[\alpha-\epsilon,\alpha+\epsilon]-\{\alpha\}$, dicho ϵ existe pues tanto $P_n(x)$ como sus derivadas son polinomios no idénticamente nulos. Entonces se tiene $\forall s=0,1,\ldots,k-1$ que si $P_n^{s)}(\alpha-\epsilon)>0$ necesariamente $P_n^{s+1)}(\alpha-\epsilon)<0$ y si $P_n^{s)}(\alpha-\epsilon)<0$, entonces $P_n^{s+1)}(\alpha-\epsilon)>0$. Por otra parte, $\forall s=0,1,\ldots,k-1$ también se tiene que el signo de $P_n^{s)}(\alpha+\epsilon)$ y de $P_n^{s+1)}(\alpha+\epsilon)$ coinciden, por tanto $N(\alpha-\epsilon)-N(\alpha+\epsilon)=k$.

Si α no es raíz de $P_n(x)$ pero es raíz de orden k de $P_n^{s)}(x)$ con $s+k<n$ se tiene que $P_n^{k-1)}(\alpha)\neq 0$, $P_n^{k)}(\alpha)=\cdots=P_n^{k+s-1)}(\alpha)=0$, $P_n^{k+s)}(\alpha)\neq 0$. Por otra parte, $\exists\epsilon>0$ de modo que ningún polinomio de la secuencia se anula en el conjunto $[\alpha-\epsilon,\alpha+\epsilon]-\{\alpha\}$, y puesto que $P_n^{k-1)}(x),P_n^{k+s)}(x)$ no cambian de signo en $[\alpha-\epsilon,\alpha+\epsilon]$ se tiene que la variación de signos en la secuencia será nulo o será par.

Para finalizar ,cabe la posibilidad de que en $x=a$ ó en $x=b$ no se anule $P_n(x)$ pero sí lo haga alguna de sus derivadas. Entonces se tiene que $\exists\epsilon>0$ de modo que los polinomios de la secuencia no se anulan en $]a,a+\epsilon]$ ni en $[b-\epsilon,b[$ por tanto el número $N(a,b)$ de raíces de $P_n(x)$ cumple $N(a,b)=N(a+\epsilon,b-\epsilon)=N(a+\epsilon)-N(b-\epsilon)-2k=\overline{N}(a)-\underline{N}(b)-2k$ con , $k=0,1,\ldots E\left[\frac{\underline{N}(a)-\overline{N}(b)}{2}\right]$. ▼

Ejemplo 5.5.7

Determinar a partir del teorema de Boundan-Fourier el número de raíces del polinomio $P_n(x)=x^4+2x^3-3x^2-4x+4$ en el intervalo $[\frac{1}{2},\frac{3}{2}]$.

Sea la secuencia $P_n(x),P'_n(x),P''_n(x),P'''_n(x),P_n^{iv)}(x),P_n^{v)}(x)$ dicha secuencia viene dada de modo explícito por

$$x^4+2x^3-3x^2-4x+4;\ 4x^3+6x^2-6x+4;\ 12x^2+12x-6;\ 24x+12,24$$

Así, para $x=\frac{1}{2}$ se tiene la secuencia $\frac{25}{16},-5,3,24,24$, y para $x=\frac{3}{2}$ la secuencia resulta $\frac{49}{16},14,39,48,24$ y, por tanto, $\overline{N}\left(\frac{1}{2}\right)=2$ y $\underline{N}\left(\frac{3}{2}\right)=0$.

Por tanto, dicho polinomio tiene bien dos raíces, bien ninguna en dicho intervalo. ♦

Teorema 5.5.20 (Regla de los signos de Descartes)

Sea $P_n(x)=a_nx_n+a_{n-1}x^n-1+\cdots+a_0$ un polinomio de grado $n\geq 1$ con coeficientes reales. Entonces el número de raíces positivas de $P_n(x)$ donde cada raíz cuenta según su multiplicidad viene dado por $N-2k$ donde N es el número de cambios de signo de la secuencia $(a_0,a_1,\ldots,a_n)$ y k un número entero comprendido entre 0 y $E\left[\frac{N}{2}\right]$.

Dem:

El número de raíces reales positivas de $P_n(x)$ es, según el teorema de Boundan-Fourier, $N(0,+\infty)=\underline{N}(0)-\overline{N}(\infty)-2k$. Pero, por una parte

$$(P_n(0),P'_n(0),\ldots,P_n^{n)}(0))=(a_0,a_1,\ldots,k!a_k,\ldots,n!a_n)$$

cuyos signos coinciden con los de $(a_0, a_1, \ldots, a_n)$, por tanto $\underline{N}(0) = N$

Por otra parte, se tiene que $\overline{N}(+\infty) = 0$ pues

$$(P_n(+\infty), P_n'(+\infty), \ldots, P_n^{n)}(+\infty)) = (+\infty, +\infty, \ldots, +\infty),$$

así, $N(0, +\infty) = N - 2k$ con $k = 0, 1, \ldots, E\left[\frac{N}{2}\right]$. ▼

Teorema 5.5.21 (Teorema de Hua)
Sea $P_n(x) = a_n x_n + a_{n-1}x^{n-1} + \cdots + a_0$ un polinomio de grado $n \geq 1$. Entonces para que todas sus raíces sean reales es necesario que $a_i^2 > a_{i-1}a_{i+1},\ i = 1, \ldots, n-1$.

Ejemplo 5.5.8
¿Pueden tener los polinomios $P(x) = x^4 + 2x^3 - 3x^2 - 4x + 4$ y $Q(x) = x^4 + x^3 + 2x^2 + x + 1$ todas sus raíces reales?

En el primer caso se tiene que

$$\begin{aligned}
a_3^2 &= \ \ 2,\ a_2 \cdot a_4 = \ \ 3 \rightarrow a_3^2 > a_2 \cdot a_4 \\
a_2^2 &= \ \ 9,\ a_1 \cdot a_3 = \ \ 8 \rightarrow a_2^2 > a_1 \cdot a_3 \\
a_1^2 &= 16,\ a_0 \cdot a_2 = 12 \rightarrow a_1^2 > a_0 \cdot a_2.
\end{aligned}$$

Por tanto, dicho polinomio sí que puede tener todas sus raíces reales.

En el otro caso se tiene que $b_3^2 = 1,\ b_2 \cdot b_4 = 2 \rightarrow b_3^2 \leq b_2 \cdot b_4$ y, por tanto, no se satisface $b_3^2 > b_2 \cdot b_4$ de donde no todas las raíces de la ecuación $Q(z) = 0$ pueden ser reales. ♦

5.5.3. Método de Bairstow

El método de Bairstow pretende resolver una ecuación polinómica $P_n(x) = 0$ para $n > 2$ mediante su descomposición como producto de un factor cuadrático y de un polinomio $Q(x)$ de grado $n - 2$; de este modo se procede a resolver la ecuación resultante de igualar a cero dicho factor cuadrático y reducir el problema, a uno de una ecuación polinómica cuyo grado es menor en dos unidades al polinomio inicial.
Sea un polinomio $P_n(z) = \sum\limits_{i=0}^{n} a_i z^i$ y sea el factor cuadrático $z^2 - uz - v$, entonces podemos escribir

$$P_n(z) = (z^2 - uz - v)Q_n(z) + r(z),$$

donde $r(z)$ es un polinomio de grado menor que dos.
Por conveniencia, y sin pérdida de generalidad, escribiremos $Q(z)$ y $r(z)$ como

$$Q(z) = \sum_{i=2}^{n} b_i z^{i-2}, \quad r(z) = b_1(z - u) + b_0.$$

Efectuando operaciones resulta

$$(z^2 - uz - v)Q_n(z) + r(z) = b_n z^n + (b_{n-1} - ub_n)z^{n-1} + \sum_{i=0}^{n-2}(b_i - ub_{i+1} - vb_{i+2})z^i,$$

definiendo $b_{n+1} = b_{n+2} = 0$ se tiene

$$(z^2 - uz - v)Q_n(z) + r(z) = \sum_{i=0}^{n} (b_i - ub_{i+1} - vb_{i+2})z^i,$$

y por tanto:

$$b_k = a_k + ub_{k+1} + vb_{k+2}, \qquad \forall k = n, n-1, .., 2, 1.$$

Para que $P_n(x)$ sea múltiplo de $z^2 - uz - v$ es necesario y suficiente que $b_0 = b_1 = 0$, por tanto, podemos plantear el problema de encontrar valores (u, v) de modo que $P_n(x)$ sea múltiplo de $z^2 - uz - v$

$$\begin{array}{c} b_0(u, v) = 0 \\ b_1(u, v) = 0 \end{array}$$

lo que constituye un sistema de ecuaciones no lineales que resolveremos utilizando el método de Newton Ralphson. Así, si (u_r, v_r) es una aproximación a la raíz del sistema $b_0(u, v) = b_1(u, v) = 0$ se tiene que la siguiente aproximación (u_{r+1}, v_{r+1}) viene dada por

$$\begin{array}{c} u_{r+1} = u_r + \Delta u_r \\ v_{r+1} = v_r + \Delta v_r \end{array} \quad \text{donde} \begin{bmatrix} \frac{\partial b_0}{\partial u} & \frac{\partial b_0}{\partial v} \\ \frac{\partial b_1}{\partial u} & \frac{\partial b_1}{\partial v} \end{bmatrix}_{(u_r, v_r)} \begin{bmatrix} \Delta u_r \\ \Delta v_r \end{bmatrix} = \begin{bmatrix} -b_0(u_r, v_r) \\ -b_1(u_r, v_r) \end{bmatrix}.$$

Definiendo $c_r = \frac{\partial b_r}{\partial u}$ y $d_r = \frac{\partial b_{r-1}}{\partial v}$, haciendo $c_n = c_{n+1} = 0$, $d_n = d_{n+1} = 0$ y derivando la recurrencia que nos proporcionan los b_k se tiene

$$\begin{array}{c} c_k = b_{k+1} + uc_{k+1} + vc_{k+2} \\ d_k = b_{k+1} + ud_{k+1} + vd_{k+2} \end{array}, \quad k = n-1, n-2, \ldots, 2, 1.$$

y puesto que coinciden las recurrencias y las condiciones iniciales, se tiene que $c_k = d_k$.
Finalmente, se obtiene que $(\Delta u_r, \Delta v_r)$ satisfacen

$$\begin{bmatrix} c_0 & c_1 \\ c_1 & c_2 \end{bmatrix} \begin{bmatrix} \Delta u_r \\ \Delta v_r \end{bmatrix} = \begin{bmatrix} -b_0 \\ -b_1 \end{bmatrix}.$$

Ejemplo 5.5.9
Resuélvase la ecuación $z^4 - 4z^3 + 7z^2 - 5z - 2$ mediante el método de Bairstow.

El método de Bairstow se basa en la obtención de factores cuadráticos de la forma $z^2 - uz - v$. Para ello, dados unos valores iniciales de (u, v) se procede a determinar los coeficientes b_k definidos como

$$b_k = a_k + u\, b_{k+1} + v\, b_{k+2}, \qquad b_{n+2} = b_{n+1} = 0,$$

a partir de los cuales se determinan los valores $c_k = \frac{\partial b_k}{\partial u}$ y $d_k = \frac{\partial b_{k-1}}{\partial v}$, valores que se determinan como:

$$c_k = b_{k+1} + u\, c_{k+1} + v\, c_{k+2}, \qquad c_{n+1} = c_n = 0$$

$$c_k = b_{k+1} + u\, c_k + 1 + v\, c_k + 2, \qquad c_{n+1} = c_n = 0$$

los valores d_k coinciden con los de c_k.
A continuación, se procede a resolver el sistema

$$\begin{bmatrix} c_0 & c_1 \\ c_1 & c_2 \end{bmatrix} \begin{bmatrix} \Delta u_0 \\ \Delta v_0 \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \end{bmatrix},$$

El método se aplica iterativamente hasta que $||(\Delta u, \Delta v)|| < \mathrm{err}$. Así, para $(u_0, v) = (0, 0)$ resulta:

k	4	3	2	1	0
b_k	1	-4	7	-5	-2
c_k	0	1	-4	7	-5

A continuación, se procede a resolver el sistema:

$$\begin{bmatrix} -5 & 7 \\ 7 & -4 \end{bmatrix} \begin{bmatrix} \Delta u_0 \\ \Delta v_0 \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \end{bmatrix}, \qquad \begin{bmatrix} \Delta u_0 \\ \Delta v_0 \end{bmatrix} = \begin{bmatrix} 1.483 \\ 1.345 \end{bmatrix}.$$

Repitiendo los cálculos para $(u_1, v_1) = (1.483, 1.345)$ donde $u_1 = u_0 + \Delta u_0$, $v_1 = v_0 + \Delta v_1$, se tiene

k	4	3	2	1	0
b_k	1.000	-2.517	4.612	-1.546	1.910
c_k	0.00	1.000	-1.034	4.423	3.621

$$\begin{bmatrix} 3.621 & 4.423 \\ 4.423 & -1.034 \end{bmatrix} \begin{bmatrix} \Delta u_1 \\ \Delta v_1 \end{bmatrix} = \begin{bmatrix} -0.910 \\ -1.546 \end{bmatrix}, \qquad \begin{bmatrix} \Delta u_1 \\ \Delta v_1 \end{bmatrix} = \begin{bmatrix} 0.209 \\ -0.603 \end{bmatrix},$$

Por tanto, $(u_2, v_2) = (1.691, 0.742)$. Iterando de nuevo se tiene:

k	4	3	2	1	0
b_k	1.000	-2.309	3.837	-0.223	0.471
c_k	0.00	1.000	-0.617	3.536	5.299

$$\begin{bmatrix} 5.299 & 3.536 \\ 3.536 & -0.617 \end{bmatrix} \begin{bmatrix} \Delta u_2 \\ \Delta v_2 \end{bmatrix} = \begin{bmatrix} -0.471 \\ -0.223 \end{bmatrix}, \qquad \begin{bmatrix} \Delta u_2 \\ \Delta v_2 \end{bmatrix} = \begin{bmatrix} 0.031 \\ -0.181 \end{bmatrix}.$$

Ahora se tiene que $(u_3, v_3) = (1.722, 0.561)$. Iterando de nuevo se obtiene:

k	4	3	2	1	0
b_k	1.000	-2.277	3.638	-0.010	0.026
c_k	0.000	1.000	-0.554	3.245	5.270

$$\begin{bmatrix} 5.270 & 3.245 \\ 3.245 & -0.554 \end{bmatrix} \begin{bmatrix} \Delta u_3 \\ \Delta v_3 \end{bmatrix} = \begin{bmatrix} -0.026 \\ 0.010 \end{bmatrix}, \qquad \begin{bmatrix} \Delta u_3 \\ \Delta v_3 \end{bmatrix} = \begin{bmatrix} 0.001 \\ -0.010 \end{bmatrix}.$$

Por tanto, $(u_4, v_4) = (1.723, 0.551)$. Iterando de nuevo se tiene:

k	4	3	2	1	0
b_k	1.000	-2.276	3.627	0.000	0.000
c_k	0.00	1.000	-0.551	3.228	5.262

$$\begin{bmatrix} 5.262 & 3.228 \\ 3.228 & -0.551 \end{bmatrix} \begin{bmatrix} \Delta u_4 \\ \Delta v_4 \end{bmatrix} = \begin{bmatrix} 0.000 \\ 0.000 \end{bmatrix}, \qquad \begin{bmatrix} \Delta u_4 \\ \Delta v_4 \end{bmatrix} = \begin{bmatrix} 0.000 \\ 0.000 \end{bmatrix}$$

con lo que termina el proceso. El polinomio inicial queda factorizado como:

$$z^4 - 4z^3 + 7z^2 - 5z - 2 = \left(z^2 - 1.723z - 0.551\right)\left(1.000z^2 - 2.276z + 3.627\right).$$

Resolviendo las dos ecuaciones de segundo grado se tiene que las raíces son:

$$z_1 = -0.276, \qquad z_2 = 1.999, \qquad z_3 = 1.138 + 1.527i, \qquad z_4 = 1.138 + 1.527i.$$

♦

Tema 6

Sistemas lineales de ecuaciones

6.1. Introducción

Uno de los tópicos más importantes de los métodos numéricos consiste en el estudio de métodos eficaces para la resolución de sistemas lineales de ecuaciones.

En el presente capítulo un sistema de ecuaciones lineales lo representaremos como $A\overrightarrow{x} = \overrightarrow{b}$, donde A es una matriz real o compleja regular $n \times n$ y que denominaremos matriz del sistema, $\overrightarrow{b}$ es un vector de $\mathbb{R}^n$ llamado vector de términos independientes y $\overrightarrow{x}$ es un vector de $\mathbb{R}^n$ que representa la solución del sistema de ecuaciones.
De forma completa, el sistema anterior se representa como:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & \cdots & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & \cdots & \cdots & a_{2,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n,1} & a_{n,2} & \cdots & \cdots & \cdots & a_{n,n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ \cdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ \cdots \\ b_n \end{bmatrix}.$$

Los métodos para la resolución de tales sistemas se clasifican como métodos directos, cuando nos proporcionan la solución exacta del sistema en un número finito de operaciones, y métodos iterativos, si están basados en un esquema de aproximaciones sucesivas de convergentes a la solución.

6.2. Métodos directos

Para abordar el estudio de los métodos directos estudiaremos en primer lugar unos tipos especiales de sistemas de ecuaciones cuya solución exacta se puede escribir de modo inmediato, tales sistemas son los llamados sistemas diagonales $D\overrightarrow{x} = \overrightarrow{b}$ de modo que $d_{i,j} \begin{cases} = 0 & \text{si } i \neq j \\ \neq 0 & \text{si } i = j \end{cases}$.
Este tipo de sistema se resuelve de modo inmediato mediante

$$x_i = \frac{b_i}{d_{i,i}}, \quad i = 1, \ldots, n.$$

Otra clase de sistemas que se resuelve fácilmente son los llamados sistemas triangulares inferiores, estos sistemas se suelen representar como $L\overrightarrow{x} = \overrightarrow{b}$ donde L es una matriz triangular inferior, y puesto que debe ser no singular, se tiene que todos los elementos de la diagonal principal son distintos de cero. Así, se tiene el sistema

$$\begin{bmatrix} \ell_{1,1} & 0 & 0 & \cdots & \cdots & 0 \\ \ell_{2,1} & \ell_{2,2} & 0 & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ \ell_{2,1} & \ell_{2,2} & 0 & \cdots & \ell_{n-1,n-1} & 0 \\ \ell_{2,1} & \ell_{2,2} & 0 & \cdots & \ell_{n,n-1} & \ell n, n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_{n-1} \\ b_n \end{bmatrix}$$

La solución de este sistema viene dada mediante el algoritmo

$$x_i = \begin{cases} \frac{b_1}{\ell_{1,1}} & i = 1 \\ \frac{b_i - \sum\limits_{j=1}^{i-1} \ell_{i,j} x_j}{\ell_{i,i}} & i = 2, \ldots, n \end{cases}.$$

El caso de un sistema triangular superior $U\overrightarrow{x} = \overrightarrow{b}$ donde U es una matriz triangular superior y los elementos de la diagonal son no nulos:

$$\begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & \cdots & u_{1,n-1} & u_{1,n} \\ 0 & u_{2,2} & \cdots & \cdots & u_{2,n-1} & u_{2,n} \\ 0 & 0 & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \cdots & u_{n-1,n-1} & u_{n-1,n} \\ 0 & 0 & \cdots & \cdots & 0 & u_{n,n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_{n-1} \\ b_n \end{bmatrix}.$$

La solución de este sistema viene dada mediante el algoritmo

$$x_i = \begin{cases} \frac{b_n}{u_{n,n}} & i = n \\ \frac{b_i - \sum\limits_{j=i+1}^{n} u_{i,j} x_j}{u_{i,i}} & i = n-1, n-2, \ldots, 1 \end{cases}$$

Los métodos directos tratan de reducir el problema a uno o varios de estos casos, bien mediante la transformación del sistema inicial en uno de estos a base de efectuar adecuadas combinaciones lineales, lo que constituye los llamados métodos tipo Gauss, bien mediante una adecuada descomposición de la matriz A del sistema como producto de matrices de tipo L, D, U, reduciendo de este modo el problema a la resolución secuencial de los problemas elementales anteriormente abordados, lo que constituye los llamados métodos de descomposición.

6.2.1. Métodos gaussianos

Como es bien conocido, el método de Gauss consiste en la sustitución de un sistema de ecuaciones lineales por otro equivalente triangular superior, lo que se consigue teniendo en cuenta que, dado un sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$, las siguientes operaciones transforman el sistema en otro equivalente (entendiendo por equivalente el tener la misma solución) al primero:

1. El intercambio de dos ecuaciones entre sí.

2. La sustitución de una ecuación por la combinación lineal de ecuaciones en la que aparezca la ecuación sustituida con coeficiente no nulo.

Con estas reglas se procede como sigue: dado el sistema lineal

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & \cdots & a_{2,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n,1} & a_{n,2} & \cdots & \cdots & a_{n,n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ \cdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ \cdots \\ b_n \end{bmatrix},$$

se reescribe como

$$\begin{bmatrix} a^{(1)}_{1,1} & a^{(1)}_{1,2} & \cdots & \cdots & a^{(1)}_{1,n} \\ a^{(1)}_{2,1} & a^{(1)}_{2,2} & \cdots & \cdots & a^{(1)}_{2,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a^{(1)}_{n,1} & a^{(1)}_{n,2} & \cdots & \cdots & a^{(1)}_{n,n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ \cdots \\ x_n \end{bmatrix} = \begin{bmatrix} b^{(1)}_1 \\ b^{(1)}_2 \\ \cdots \\ \cdots \\ b^{(1)}_n \end{bmatrix}.$$

donde $a^{(1)}_{i,j} = a_{i,j}$ y $b^{(1)}_i = b_1$ si $a_{1,1} \neq 0$. En el caso en que $a_{1,1} = 0$ sea i_1 la primera ecuación para la que $a_{i_1,1} \neq 0$, entonces se define

$$a^{(1)}_{i,j} = \begin{cases} a_{i,j} & i \neq 1, i_1 \\ a_{i_1,j} & i = 1 \\ a_{1,j} & i = i_1 \end{cases} \qquad b^{(1)}_i = \begin{cases} b_i & \text{si } i \neq 1, i_1 \\ b_{i_1} & i = 1 \\ b_1 & i = i_1 \end{cases} \qquad i, j = 1, \ldots, n,$$

a partir de lo cual se procede a anular todos los elementos de la primera columna salvo el primero, utilizando la transformación:

$$\text{Fila } i\text{-ésima} = \text{ Fila } i\text{-ésima} - \frac{a^{(1)}_{i,1}}{a^{(1)}_{1,1}} \text{ Fila primera.}$$

esto es

$$\alpha^{(1)}_{i,j} = a_{i,j} - \frac{a_{i,1}}{a_{1,1}} a_{1,j} \qquad \beta^{(1)}_i = b_i - \frac{a_{i,1}}{a_{1,1}} b_1 \qquad i = 2, \ldots, n$$

Si $\alpha^{(1)}_{2,2} \neq 0, \forall i, j = 2, \ldots, n$ se realiza la transformación $a^{(2)}_{i,j} = \alpha^{(1)}_{i,j}$ y $b^{(2)}_i = \beta^{(1)}_i$. Si $\alpha^{(1)}_{2,2} = 0$, sea i_2 el primer $i \in \{2, 3, \ldots, n\}$ para el que $\alpha^{(1)}_{i_2,2} \neq 0$ en cuyo caso se define:

$$a^{(2)}_{i,j} = \begin{cases} \alpha^{(1)}_{i,j} & i \neq 2, i_2 \\ \alpha^{(1)}_{i_1,j} & i = 1 \\ \alpha^{(1)}_{1,j} & i = i_2 \end{cases}, \quad b^{(2)}_i = \begin{cases} \beta^{(1)}_i & i \neq 2, i_2 \\ \beta^{(1)}_{i_2} & i = 1 \\ \beta^{(1)}_2 & i = i_2 \end{cases}, \quad i, j = 2, \ldots, n,$$

con lo que el sistema se puede escribir en la forma

$$\begin{bmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & a_{1,3}^{(1)} & \cdots & \cdots & a_{1,n}^{(1)} \\ 0 & a_{2,2}^{(2)} & a_{2,3}^{(2)} & \cdots & \cdots & a_{2,n}^{(2)} \\ 0 & a_{3,2}^{(2)} & a_{3,3}^{(2)} & \cdots & \cdots & a_{3,n}^{(2)} \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & a_{n,2}^{(2)} & a_{n,3}^{(2)} & \cdots & \cdots & a_{n,n}^{(2)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(2)} \\ \cdots \\ b_n^{(2)} \end{bmatrix},$$

donde $a_{2,2}^{(2)} \neq 0$ Sea el sistema:

$$\begin{bmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & a_{1,3}^{(1)} & \cdots & a_{1,k}^{(1)} & \cdots & a_{1,n}^{(1)} \\ 0 & a_{2,2}^{(2)} & a_{2,3}^{(2)} & \cdots & a_{2,k}^{(2)} & \cdots & a_{2,n}^{(2)} \\ 0 & 0 & a_{3,k}^{(k)} & \cdots & a_{3,k}^{(3)} & \cdots & a_{3,n}^{(2)} \\ 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & a_{k,k}^{(k)} & \cdots & a_{k,n}^{(k)} \\ 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & a_{n,k}^{(k)} & \cdots & a_{n,n}^{(k)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdots \\ x_k \\ \cdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ \cdots \\ b_k^{(k)} \\ \cdots \\ b_n^{(k)} \end{bmatrix}$$

donde $a_{k,k}^{(k)} \neq 0$. Aplicando el proceso de eliminación gaussiana a la k-ésima columna se tiene:

$$\text{Fila } i\text{-ésima} = \text{ Fila } i\text{-ésima} - a_{i,k}^{(k)}/a_{k,k}^{(k)} \text{ Fila } k\text{-ésima}$$

esto es

$$\alpha_{i,j}^{(k)} = a_{i,j} - \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}} a_{k,j}^{(k)} \qquad \beta_i^{(k)} = b_i^{(k)} - \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}} b_k^{(k)} \qquad i = k+1, \ldots, n$$

Si $\alpha_{k,k}^{(k)} \neq 0$, $\forall i, j = k, \ldots, n$ se realiza la transformación $a_{i,j}^{(k+1)} = \alpha_{i,j}^{(k)}$ y $b_i^{(k+1)} = \beta_i^{(k)}$.

Si $\alpha_{k,k}^{(k)} = 0$, sea i_k el primer $i \in \{k, \ldots, n\}$ para el que $\alpha_{i_k,k}^{(k)} \neq 0$ en cuyo se define:

$$a_{i,j}^{(k+1)} = \begin{cases} \alpha_{i,j}^{(k)} & i \neq k, i_k \\ \alpha_{i_k,j}^{(k)} & i = k \\ \alpha_{k,j}^{(k)} & i = i_k \end{cases} \qquad b_i^{(k+1)} = \begin{cases} \beta_i^{(k)} & i \neq k, i_k \\ \beta_{i_k}^{(k)} & i = k \\ \beta_k^{(k)} & i = i_k \end{cases} \qquad i, j \in \{k, \ldots, n\}$$

con lo que el sistema se puede reescribir de forma

$$\begin{bmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & a_{1,3}^{(1)} & \cdots & a_{1,k}^{(1)} & \cdots & a_{1,n}^{(1)} \\ 0 & a_{2,2}^{(2)} & a_{2,3}^{(2)} & \cdots & a_{2,k}^{(2)} & \cdots & a_{2,n}^{(2)} \\ 0 & 0 & a_{3,k}^{(3)} & \cdots & a_{3,k}^{(3)} & \cdots & a_{3,n}^{(2)} \\ 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & a_{k,k}^{(k)} & \cdots & a_{k,n}^{(k)} \\ 0 & 0 & 0 & 0 & 0 & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & a_{n,n}^{(n)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdots \\ x_k \\ \cdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ \cdots \\ b_k^{(k)} \\ \cdots \\ b_n^{(k)} \end{bmatrix}$$

Con estas tranformaciones, el sistema se reduce a uno equivalente triangular superior en el que todos los elementos de la diagonal son no nulos, el cual se resuelve de modo inmediato, según procedimiento estudiado anteriormente.

Ejemplo 6.2.1
Resolver mediante el método de Gauss el sistema de ecuaciones dado por:

$$\begin{bmatrix} 2 & 1 & -4 & 11 \\ 2 & 1 & 7 & 4 \\ 1 & -3 & -2 & 8 \\ -4 & 3 & -5 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 2 \\ -3 \\ 4 \\ 5 \end{bmatrix}$$

En primer lugar, escribimos el sistema en la forma

$$\left[\begin{array}{rrrr|r} 2 & 1 & -4 & 11 & 2 \\ 2 & 1 & 7 & 4 & -3 \\ 1 & -3 & -2 & 8 & 4 \\ -4 & 3 & -5 & 0 & 5 \end{array}\right],$$

siendo $[A|b]$ la matriz cuya primera parte, $[A|$, representa la matriz A del sistema, y cuya segunda parte, $|b]$ representa el vector de términos independientes.

Puesto que $a_{1,1}$ es distinto de cero utilizamos como pivote dicho elemento, con lo que haciendo ceros en la primera columna mediante el algoritmo de eliminación gaussiana, se obtiene:

$$\left[\begin{array}{rrrr|r} 2 & 1 & -4 & 11 & 2 \\ 0 & 0 & 11 & -7 & -5 \\ 0 & -\frac{7}{2} & 0 & \frac{5}{2} & 3 \\ 0 & 5 & -13 & 22 & 9 \end{array}\right].$$

En esta tabla el elemento $a_{2,2} = 0$, por lo que se debe permutar las filas segunda y tercera

$$\left[\begin{array}{rrrr|r} 2 & 1 & -4 & 11 & 2 \\ 0 & -\frac{7}{2} & 0 & \frac{5}{2} & 3 \\ 0 & 0 & 11 & -7 & -5 \\ 0 & 5 & -13 & 22 & 9 \end{array}\right].$$

Aplicando a continuación el algoritmo de eliminación gaussiana, se obtiene:

$$\left[\begin{array}{rrrr|r} 2 & 1 & -4 & 11 & 2 \\ 0 & -\frac{7}{2} & 0 & \frac{5}{2} & 3 \\ 0 & 0 & 11 & -7 & -5 \\ 0 & 0 & -13 & \frac{179}{7} & \frac{93}{7} \end{array}\right],$$

de donde finalmente, se tiene

$$\left[\begin{array}{cccc|c} 2 & 1 & -4 & 11 & 2 \\ 0 & -\frac{7}{2} & 0 & \frac{5}{2} & 3 \\ 0 & 0 & 11 & -7 & -5 \\ 0 & 0 & 0 & \frac{1332}{77} & \frac{568}{77} \end{array}\right].$$

Resolviendo el sistema triangular superior, se tiene $x_4 = \frac{142}{333}$, $x_3 = -\frac{61}{333}$, $x_2 = -\frac{184}{333}$, $x_1 = -\frac{478}{333}$.

♦

6.2.2. Método del pivote total

En el algoritmo de Gauss para la resolución de sistemas de ecuaciones, en la i-ésima iteración se procede a dividir la i-ésima fila por el elemento $a_{i,i}$. En muchos casos, los coeficientes de un sistema de ecuaciones no son valores exactos sino que cada valor numérico esta afectado por un error (de medida, de truncamiento, etc) cuyo valor viene dado por ϵ. Sea $\alpha_{i,j}$ el valor exacto de los coeficientes y $a_{i,j}$ aproximaciones con error menor que ϵ. Para evalúar el error cometido al aproximar $\frac{\alpha}{\beta}$ por $\frac{a}{b}$, donde $\alpha = a + \triangle a$, $\beta = b + \triangle b$, se procede, en primer lugar, a desarrollar $\frac{\alpha}{\beta} = \frac{a+\Delta a}{b+\Delta b}$ en primer orden respecto a $\triangle a$, $\triangle b$. Así, se obtiene

$$\frac{a+\Delta a}{b+\Delta b} = \frac{a}{b} + \frac{\Delta a}{b} - \frac{a\Delta b}{b^2}.$$

Por tanto, en primer orden

$$\left|E\left[\frac{a}{b}\right]\right| = \left|\frac{\Delta a}{b} - \frac{a\Delta b}{b^2}\right| = \frac{|b-a|\epsilon}{b^2}.$$

En esta expresión aparece b en el denominador, por lo que su valor disminuye cuando aumenta $|b|$.

$$\alpha_{i,j}^{(1)} = a_{i,j} - \frac{a_{i,1}}{a_{1,1}}a_{1,j} \qquad \beta_i^{(1)} = b_i - \frac{a_{i,1}}{a_{1,1}}b_1 \qquad i = 2,\ldots,n.$$

Con objeto de minimizar el error, el método del pivote total propone una variación con respecto al método de Gauss, consistente en elegir como pivote en cada iteración el mayor elemento en valor absoluto de la submatriz $i \in \{k,\ldots,n\}$, $j \in \{k,\ldots,n\}$. Si dicho elemento es el que ocupa el lugar (i_k, j_k) se deberán intercambiar las filas k-ésima y i_k-ésima y las columnas k-ésima y j_k-ésima.
Como es conocido el intercambio de filas (cambiando también los respectivos términos independientes) conduce a un sistema equivalente, no siendo así en el caso de intercambio de columnas, pues cada una de ellas está ligada a una variable x_j, por lo que si se intercambian las columna r y la s, el vector de solución debe ser cambiado por $(x_{\sigma(1)}, x_{\sigma(2)},\ldots,x_{\sigma(n)})^t$ donde σ es la permutación de n elementos $\sigma = (1,\ldots,r-1,s,r+1,\ldots,s-1,r,r+1,\ldots n)$ y donde ^{t} indica transposición.

Recuérdese que una permutación σ de n elementos es una aplicación biyectiva de $\{1,2,\ldots,n\}$ en $\{1,2,\ldots,n\}$. Para denotar una permutación de n elementos, se ha utilizado la notación $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ que indica que la imagen de i es σ_i, (es decir, el lugar al que mueve σ al elemento i), lo que se lee como el i al σ_i. El conjunto de todas las permutaciones de n elementos se denota por S_n y tiene una estructura de grupo con respecto a la composición de aplicaciones.
Sea el sistema lineal $A\overrightarrow{x} = \overrightarrow{b}$ donde A es una matriz regular de tamaño $n \times n$ y $\overrightarrow{x}, \overrightarrow{b} \in \mathbb{R}^n$ dado por:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & \cdots & a_{2,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n,1} & a_{n,2} & \cdots & \cdots & a_{n,n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_n \end{bmatrix},$$

Sea (i_1, j_1) de modo que $|a_{i_1,j_1}| = \max\{|a_{i,j}|\,; i,j = 1,2,\ldots n\}$, entonces el sistema se reescribe como

$$\begin{bmatrix} a^{(1)}_{1,1} & a^{(1)}_{1,2} & \cdots & \cdots & a^{(1)}_{1,n} \\ a^{(1)}_{1,1} & a^{(1)}_{1,2} & \cdots & \cdots & a^{(1)}_{1,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a^{(1)}_{1,1} & a^{(1)}_{1,2} & \cdots & \cdots & a^{(1)}_{1,n} \end{bmatrix} \begin{bmatrix} x_{\sigma_1(1)} \\ x_{\sigma_1(2)} \\ \cdots \\ \cdots \\ x_{\sigma_1(n)} \end{bmatrix} = \begin{bmatrix} b^{(1)}_1 \\ b^{(1)}_2 \\ \cdots \\ \cdots \\ b^{(1)}_n \end{bmatrix},$$

donde

$$a^{(1)}_{i,j} = \begin{cases} a_{i,j} & i \neq 1, i_1, j \neq 1, j_1 \\ a_{i_1,j} & i = 1, \quad j \neq j_1 \\ a_{1,j} & i = i_1, \quad j \neq j_1 \\ a_{i,j_1} & i \neq 1, i_1, \quad j = 1 \\ a_{i,1} & i \neq 1, i_1, \quad j = j_1 \\ a_{i_1,j_1} & i = 1, \quad j = 1 \\ a_{1,1} & i = i_1, \quad j = j_1 \end{cases}, \quad b^{(1)}_i = \begin{cases} b_i & \text{si } i \neq 1, i_1 \\ b_{i_1} & i = 1 \\ b_1 & i = i_1 \end{cases}, \quad i,j = 1,\ldots,n,$$

siendo σ_1 la permutación $(j_1, 2, 3, \ldots, j_1 - 1, 1, j_1 + 1, \ldots n)$. Nótese que la definición de los $a^{(1)}_{i,j}$ engloba los casos en que $i_1 = 1$ o $j_1 = 1$.
A continuación se inicia el proceso de eliminación gaussiana, con lo que se obtiene

$$\alpha^{(1)}_{i,j} = \begin{cases} a_{1,j} & i = 1 \\ a_{i,j} - \frac{a_{i,1}}{a_{1,1}} a_{1,j} & i \neq 1 \end{cases} \qquad \beta^{(1)}_i = \begin{cases} b_1 & i = 1 \\ b_i - \frac{a_{i,1}}{a_{1,1}} b_1 & i \neq 1 \end{cases} \qquad i = 1,\ldots,n.$$

Sea (i_2, j_2) de modo que $|\alpha_{i_1 j_1}| = \text{máx}\{|\alpha_{i,j}|\,; i,j = 1,2,\ldots n\}$ y sea

$$a_{i,j}^{(2)} = \begin{cases} \alpha_{i,j}^{(1)} & i \neq 2, i_2, j \neq 2, j_2 \\ \alpha_{i_2,j}^{(1)} & i = 2, \quad j \neq j_2 \\ \alpha_{2,j}^{(1)} & i = i_2, \quad j \neq j_2 \\ a_{i,j_2} & i \neq 2, i_2, \quad j = 2 \\ \alpha_{i,2}^{(1)} & i \neq 2, i_2, \quad j = j_2 \\ \alpha_{i_2,j_2}^{(1)} & i = 2, \quad j = 2 \\ \alpha_{2,2}^{(1)} & i = i_2, \quad j = j_2 \end{cases} \qquad b_i^{(2)} = \begin{cases} \beta_i^{(1)} & \text{si } i \neq 2, i_2 \\ \beta_{i_2}^{(1)} & i = 2 \\ \beta_2^{(1)} & i = i_2 \end{cases} \qquad i,j = 1,\ldots,n,$$

de donde se obtiene el sistema:

$$\begin{bmatrix} a_{1,1}^{(2)} & a_{1,2}^{(2)} & a_{1,3}^{(2)} & \cdots & \cdots & a_{1,n}^{(2)} \\ 0 & a_{2,2}^{(2)} & a_{2,3}^{(2)} & \cdots & \cdots & a_{2,n}^{(2)} \\ 0 & a_{3,2}^{(2)} & a_{3,3}^{(2)} & \cdots & \cdots & a_{3,n}^{(2)} \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & a_{n,2}^{(2)} & a_{n,3}^{(2)} & \cdots & \cdots & a_{n,n}^{(2)} \end{bmatrix} \begin{bmatrix} x_{\sigma_1(1)} \\ x_{\sigma_1(2)} \\ x_{\sigma_1(3)} \\ \cdots \\ x_{\sigma_1(n)} \end{bmatrix} = \begin{bmatrix} b_1^{(2)} \\ b_2^{(2)} \\ b_3^{(2)} \\ \cdots \\ b_n^{(2)} \end{bmatrix}.$$

Sea el sistema resultante de aplicar $k-1$ veces el método

$$\begin{bmatrix} a_{1,1}^{(k)} & a_{1,2}^{(k)} & a_{1,3}^{(k)} & \cdots & a_{1,k}^{(k)} & \cdots & a_{1,n}^{(k)} \\ 0 & a_{2,2}^{(k)} & a_{2,3}^{(k)} & \cdots & a_{2,k}^{(k)} & \cdots & a_{2,n}^{(k)} \\ 0 & 0 & a_{3,k}^{(k)} & \cdots & a_{3,k}^{(k)} & \cdots & a_{3,n}^{(k)} \\ 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & a_{k,k}^{(k)} & \cdots & a_{k,n}^{(k)} \\ 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & a_{n,k}^{(k)} & \cdots & a_{n,n}^{(k)} \end{bmatrix} \begin{bmatrix} x_{\sigma_{k-1}(1)} \\ x_{\sigma_{k-1}(2)} \\ x_{\sigma_{k-1}(3)} \\ \cdots \\ x_{\sigma_{k-1}(k)} \\ \cdots \\ x_{\sigma_{k-1}(n)} \end{bmatrix} = \begin{bmatrix} b_1^{(k)} \\ b_2^{(k)} \\ b_3^{(k)} \\ \cdots \\ b_k^{(k)} \\ \cdots \\ b_n^{(k)} \end{bmatrix}.$$

Aplicando el algoritmo de eliminación gaussiana a la k-ésima columna, se tiene

$$\alpha_{i,j}^{(k)} = \begin{cases} a_{k,j}^{(k)} & i \leq k \\ a_{i,j}^{(k)} - \frac{a_{i,k}^{(k)}}{a_{k,k}(k)} a_{k,j}^{(k)} & i \neq k \end{cases} \qquad i,j = 1,\ldots,n, i \leq j$$

$$\beta_i^{(k)} = \begin{cases} b_1^{(k)} & i \leq k \\ b_i^{(k)} - \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}} b_k^{(k)} & i > k \end{cases} \qquad i,j = 1,\ldots,n, i \leq j.$$

Sea (i_k, j_k) de modo que $\left|\alpha_{i_k,j_k}^{(k)}\right| = \text{máx}\left\{\left|\alpha_{i,j}^{(k)}\right|; i,j = k, k+1, \ldots n\right\}$, y sea la permutación $\sigma_k = (1, 2, \ldots, k-1, i_k, k+1, \ldots, i_k-1, k, i_k+1, \ldots, n)\sigma_{k-1}$ (donde

el producto debe tomarse en el sentido de composición de aplicaciones), definiendo

$$a_{i,j}^{(k+1)} = \begin{cases} \alpha_{i,j}^{(k)} & i \neq k, i_k, j \neq k, j_k \\ \alpha_{i_k,j}^{(k)} & i = k, \quad j \neq j_k \\ \alpha_{k,j}^{(k)} & i = i_k, \quad j \neq j_k \\ \alpha_{i,j_k}^{(k)} & i \neq k, i_k, \quad j = k \\ \alpha_{i,k}^{(k)} & i \neq k, i_k, \quad j = j_k \\ \alpha_{i_k,j_k}^{(k)} & i = k, \quad j = k \\ \alpha_{k,k}^{(k)} & i = i_k, \quad j = j_k \end{cases} \qquad i, j = 1, \ldots, n.$$

$$b_i^{(k+1)} = \begin{cases} \beta_i^{(k)} & i \neq k, i_k \\ \beta_{i_k}^{(k)} & i = k \\ \beta_k^{(k)} & i = i_k \end{cases} \qquad i, j = 1, \ldots, n.$$

Aplicando $n-1$ veces el algoritmo se obtiene:

$$\begin{bmatrix} a_{1,1}^{(n)} & a_{1,2}^{(n)} & a_{1,2}^{(n)} & \cdots & \cdots & a_{1,n-1}^{(n)} & a_{1,n}^{(n)} \\ 0 & a_{2,2}^{(n)} & a_{2,2}^{(n)} & \cdots & \cdots & a_{2,n-1}^{(n)} & a_{2,n}^{(n)} \\ 0 & 0 & a_{3,2}^{(n)} & \cdots & \cdots & a_{3,n-1}^{(n)} & a_{3,n}^{(n)} \\ 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & a_{n-1,n-1}^{(n)} & a_{n-1,n}^{(n)} \\ 0 & 0 & 0 & 0 & 0 & 0 & a_{n,n}^{(n)} \end{bmatrix} \begin{bmatrix} x_{\sigma_{n-1}(1)} \\ x_{\sigma_{n-1}(2)} \\ x_{\sigma_{n-1}(3)} \\ \cdots \\ x_{\sigma_{n-1}(n-1)} \\ x_{\sigma_{n-1}(n)} \end{bmatrix} = \begin{bmatrix} b_1^{(n)} \\ b_2^{(n)} \\ b_3^{(n)} \\ \cdots \\ b_{n-1}^{(n)} \\ b_n^{(n)} \end{bmatrix}$$

que es un sistema triangular superior cuya solución viene dada por

$$x_{\sigma_{n-1}(i)} = \frac{b_i^k - \sum\limits_{j=i+1}^{n} a_{i,j}^{(k)} x_{\sigma_{n-1}(j)}}{a_{i,i}^{(k)}}.$$

6.2.3. Métodos de descomposición

En este epígrafe se aborda el estudio de métodos para la descomposición de matrices $A_{n\times n}$ en producto de una matriz triangular inferior L y de una matriz triangular superior U.

Sean las matrices A, L, U;

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix}, \quad L = \begin{bmatrix} \ell_{1,1} & 0 & \cdots & 0 \\ \ell_{2,1} & \ell_{2,2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \ell_{n,1} & \ell_{n,2} & \cdots & \ell_{n,n} \end{bmatrix},$$

$$U = \begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,n} \\ 0 & u_{2,2} & \cdots & u_{2,n} \\ 0 & 0 & \cdots & \cdots \\ 0 & 0 & 0 & u_{n,n} \end{bmatrix}.$$

Así, se tiene:

$$\begin{bmatrix} \ell_{1,1} & 0 & \cdots & 0 \\ \ell_{2,1} & \ell_{2,2} & \cdots & 0 \\ \dots & \cdots & \cdots & \cdots \\ \ell_{n,1} & \ell_{n,2} & \cdots & \ell_{n,n} \end{bmatrix} \begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,n} \\ 0 & u_{2,2} & \cdots & u_{2,n} \\ 0 & 0 & \cdots & \cdots \\ 0 & 0 & 0 & u_{n,n} \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix}$$

y por tanto:

$$\sum_{r=1}^{\text{mín}\{i,j\}} \ell_{i,r} u_{r,j} = a_{i,j}$$

lo que constituye un sistema de n^2 ecuaciones con $n^2 + n$ incógnitas, por lo que se debe fijar n condiciones adicionales para obtener una única solución. Entre las soluciones mas habituales se encuentra hacer $\ell_{i,i} = 1, \quad i = 1, \dots, n$ y $u_{i,i} = 1, \quad i = 1, \dots, n$, lo que da lugar a la descomposición de Doolittle en el primer caso y la descomposición de Crout en el segundo.

Las ecuaciones resultantes para la descomposición de Doolittle son las siguientes:

Para cada $i = 1, \dots, n$

$$\begin{aligned} \ell_{i,i} &= 1, \\ u_{i,j} &= a_{i,j} - \sum_{r=1}^{i-1} \ell_{i,r} u_{r,j}, \qquad \forall j = i, i+1, \dots, n, \\ \ell_{j,i} &= \frac{a_{j,i} - \sum_{r=1}^{i-1} \ell_{j,r} u_{r,i}}{u_{i,i}}, \qquad \forall j = i+1, \dots, n. \end{aligned}$$

Para el caso de la descomposición de Crout resulta para cada $i = 1, \dots, n$:

$$\begin{aligned} u_{i,i} &= 1, \\ \ell_{j,i} &= a_{j,i} - \sum_{r=1}^{i-1} \ell_{j,r} u_{r,i}, \qquad \forall j = i, \dots, n, \\ u_{i,j} &= \frac{a_{i,j} - \sum_{r=1}^{i-1} \ell_{i,r} u_{r,j}}{\ell_{i,i}}, \qquad \forall j = i+1, \dots, n. \end{aligned}$$

En el caso en que la matriz A sea real, simétrica y definida positiva, la matriz U puede tomarse como $U = L^t$ (donde ^{t} indica trasposición). En este caso, el método recibe el nombre de método de Choleski o de la raíz cuadrada, resultando en este caso las ecuaciones, para cada valor de $i = 1, \dots, n$:

$$\begin{aligned} \ell_{i,i} &= \sqrt{a_{i,j} - \sum_{r=1}^{i-1} \ell_{i,s}^2}, \\ \ell_{i,j} &= \frac{a_{i,j} - \sum_{r=1}^{i-1} \ell_{i,r} \ell_{r,j}}{\ell i, i}, \qquad j = i+1, \dots, n. \end{aligned}$$

6.2.4. Sistemas Tridiagonales

Muchos problemas del cálculo numérico conducen a una clase especial de sistemas de ecuaciones lineales llamados tridiagonales, en los cuales la matriz del sistema está formada por ceros excepto en la diagonal principal, la subdiagonal y la superdiagonal.
Es por su importancia por lo que se estudia un método directo muy eficiente para su resolución. Sea el sistema

$$\begin{bmatrix} d_1 & c_1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ a_1 & d_2 & c_2 & 0 & \cdots & 0 & 0 & 0 \\ 0 & a_2 & d_3 & c_3 & \cdots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & d_{n-2} & c_{n-2} & 0 \\ 0 & 0 & 0 & 0 & \cdots & a_{n-2} & d_{n-1} & c_{n-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 & a_{n-1} & d_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_2 \\ \cdots \\ \cdots \\ x_{n-2} \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_2 \\ \cdots \\ \cdots \\ b_{n-2} \\ b_{n-1} \\ b_n \end{bmatrix}.$$

El sistema se reduce a triangular superior mediante las transformaciones

$$\delta_1 = d_1, \quad \beta_1 = b_1,$$

$$\delta_i = d_i - \frac{a_{i-1}}{\delta_{i-1}} c_i, \quad \beta_i = b_i - \frac{a_{i-1}}{\delta_{i-1}} \beta_{i-1}, \quad \forall i = 2, \ldots, n,$$

de donde resulta el sistema:

$$\begin{bmatrix} \delta_1 & c_1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \delta_2 & c_2 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \delta_3 & c_3 & \cdots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & \delta_{n-2} & c_{n-2} & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & \delta_{n-1} & c_{n-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \delta_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_2 \\ \cdots \\ \cdots \\ x_{n-2} \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_2 \\ \cdots\cdots \\ \beta_{n-2} \\ \beta_{n-1} \\ \beta_n \end{bmatrix}.$$

cuya solución viene dada de modo inmediato por

$$x_n = \frac{\beta_n}{\delta_n}, \qquad x_i = \frac{\beta_i - \sum\limits_{j=i}^{n-1} c_j x_{j+1}}{\delta_i}, \quad i = n-1, \ldots, 2, 1.$$

6.3. Métodos iterativos

Una técnica alternativa para la resolución de sistemas de ecuaciones lineales la constituyen los llamados métodos iterativos. Estos métodos están basados en lo siguiente:
Sea el sistema de ecuaciones lineales $A\overrightarrow{x} = \overrightarrow{b}$ donde A es una matriz no singular de números reales o complejos de tamaño $n \times n$, $\overrightarrow{x}$, $\overrightarrow{b}$ vectores de $\mathbb{R}^n$ o $\mathbb{C}^n$.

Sea Q una matriz regular de tamaño $n \times n$, real o compleja según sea el sistema, a la cual llamaremos matriz de partición. Entonces, el sistema inicial resulta equivalente a

$$Q\overrightarrow{x} = (Q - A)\overrightarrow{x} + \overrightarrow{b}.$$

A partir de la cual establecemos la recurrencia

$$Q\overrightarrow{x}^{(k+1)} = (Q - A)\overrightarrow{x}^{(k)} + \overrightarrow{b}, \quad k = 0, 1, \ldots, n, \ldots$$

con lo que a partir de una solución inicial $\overrightarrow{x}^0$ se determina la sucesión $\left\{\overrightarrow{x}^{(k)}\right\}_{k=0}^{\infty}$, la cual si converge lo hace a un punto $\overrightarrow{x} = \lim\limits_{k\to\infty} \overrightarrow{x}^{(k)}$ pues, dado que las aplicaciones lineales son continuas, si N es una matriz $n \times n$, se verifica

$$\lim_{k\to\infty} N\overrightarrow{x}^{(k)} = N \lim_{k\to\infty} \overrightarrow{x}^{(k)} = N\overrightarrow{x},$$

por lo que dicho límite, si existe, satisface $Q\overrightarrow{x} = (Q - A)\overrightarrow{x} + \overrightarrow{b}$ y por tanto, el sistema original $A\overrightarrow{x} = \overrightarrow{b}$.

Una condición suficiente de convergencia para la sucesión $\left\{\overrightarrow{x}^{(k)}\right\}_{k=0}^{\infty}$ definida anteriormente es que $||M|| < 1$ donde:

$$||M|| = \text{máx}\left\{||M(\overrightarrow{x})||, \,|\, ||\overrightarrow{x}|| = 1\right\},$$

es la norma matricial subordinada a la norma de $\mathbb{R}^n$, y $M = I - Q^{-1}A$. En este caso, dado que el sistema lo podemos escribir como

$$\overrightarrow{x} = \overrightarrow{\Phi}(\overrightarrow{x}) = M\overrightarrow{x} + \overrightarrow{\beta}, \qquad \overrightarrow{\beta} = Q^{-1}\overrightarrow{b}$$

se verifica que $\forall \overrightarrow{x}, \overrightarrow{y} \in \mathbb{R}^n$,

$$\left|\left|\overrightarrow{\Phi}(\overrightarrow{x}) - \overrightarrow{\Phi}(\overrightarrow{y})\right|\right| = \left|\left|(M\overrightarrow{x} + \overrightarrow{\beta}) - (M\overrightarrow{y} - \overrightarrow{\beta})\right|\right| =$$

$$= ||(M(\overrightarrow{x} - \overrightarrow{y})|| \leq ||M||\, ||\overrightarrow{x} - \overrightarrow{y}||,$$

y dado que $||M|| < 1$, se tiene que $\overrightarrow{\Phi} : \mathbb{R}^n \to \mathbb{R}^n$ es contractiva y, por tanto, tiene un único punto fijo $\overrightarrow{x}$ al cual converge la sucesión $\left\{\overrightarrow{x}^{(k)}\right\}_{k=0}^{\infty}$ cualquiera que sea el punto de partida $\overrightarrow{x}_0$ que se tome.

A continuación, se estudian en mayor detalle tres de estos métodos.

6.3.1. Método de Richardson

Este método consiste en tomar como matriz de partición la identidad, con lo que la iteración resulta:

$$\overrightarrow{x}^{(k+1)} = (I - A)\overrightarrow{x}^{(k)} + \overrightarrow{b}, \quad \overrightarrow{x}^{(0)} = \overrightarrow{b}$$

lo que escrito de modo explícito resulta:

$$\begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ \cdots \\ \cdots \\ x_n^{(k+1)} \end{bmatrix} = \begin{bmatrix} 1 - a_{1,1} & -a_{1,2} & \cdots & \cdots & -a_{1,n} \\ -a_{2,1} & 1 - a_{2,2} & \cdots & \cdots & -a_{2,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -a_{n,1} & -a_{n,2} & \cdots & \cdots & 1 - a_{n,n} \end{bmatrix} \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ \cdots \\ \cdots \\ x_n^{(k)} \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ \cdots \\ b_n \end{bmatrix}.$$

Este método es interesante cuando la matriz del sistema es próxima a la identidad.

6.3.2. Método de Jacobi

El método de Jacobi requiere que la diagonal de la matriz A del sistema no contenga ceros. En este método se toma como matriz de partición la matriz Q dada por $q_{i,j} = a_{i,j}\delta_{i,j}$, $\forall i,j = 1,\ldots,n$ ($\delta_{i,j}$ representa la delta de Kronecher que toma valor cero si $i \neq j$ y toma valor uno cuando $i = j$). Así, el sistema resulta:

$$\begin{bmatrix} a_{1,1} & 0 & \cdots & 0 & 0 \\ 0 & a_{2,2} & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_{n-1,n-1} & 0 \\ 0 & 0 & \cdots & 0 & a_{n,n} \end{bmatrix} \begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ \cdots \\ x_{n-1}^{(k+1)} \\ x_n^{(k+1)} \end{bmatrix} =$$

$$= \begin{bmatrix} 0 & -a_{1,2} & \cdots & -a_{1,n-1} & -a_{1,n} \\ -a_{2,1} & 0 & \cdots & -a_{2,n-1} & -a_{2,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -a_{n-1,1} & -a_{n-1,2} & \cdots & 0 & -a_{n-1,n} \\ -a_{n,1} & -a_{n,2} & \cdots & -a_{n,n-1} & 0 \end{bmatrix} \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ \cdots \\ x_{n-1}^{(k)} \\ x_n^{(k)} \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_{n-1} \\ b_n \end{bmatrix}.$$

cuya solución puede obtenerse mediante:

$$\begin{array}{lllllllll} x_1^{(k+1)} & = & \cdots & -\frac{a_{1,2}}{a_{1,1}}x_2^{(k)} & -\frac{a_{1,3}}{a_{1,1}}x_3^{(k)} & \cdots & -\frac{a_{1,n-1}}{a_{1,1}}x_{n-1}^{(k)} & -\frac{a_{1,n}}{a_{1,1}}x_n^{(k)} & +\frac{b_1}{a_{1,1}} \\ x_2^{(k+1)} & = & -\frac{a_{2,1}}{a_{2,2}}x_1^{(k)} & \cdots & -\frac{a_{2,3}}{a_{2,2}}x_2^{(k)} & \cdots & -\frac{a_{2,n-1}}{a_{2,2}}x_{n-1}^{(k)} & -\frac{a_{2,n}}{a_{2,2}}x_n^{(k)} & +\frac{b_2}{a_{2,2}} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_n^{(k+1)} & = & -\frac{a_{n,1}}{a_{n,n}}x_1^{(k)} & -\frac{a_{n,2}}{a_{n,n}}x_2^{(k)} & -\frac{a_{n,3}}{a_{n,n}}x_3^{(k)} & \cdots & -\frac{a_{n,n-1}}{a_{n,n}}x_{n-1}^{(k)} & \cdots & +\frac{b_n}{a_{n,n}} \end{array}.$$

Ejemplo 6.3.1
Resolver mediante el método de Jacobi el sistema de ecuaciones:

$$\begin{bmatrix} 2 & -1 & 0 \\ 1 & 6 & -2 \\ 4 & -3 & 8 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2 \\ -4 \\ 5 \end{bmatrix}.$$

En primer lugar, dividimos cada fila por el elemento de la diagonal resultando:

$$\begin{bmatrix} 1 & -\frac{1}{2} & 0 \\ \frac{1}{6} & 1 & -\frac{1}{3} \\ \frac{1}{2} & -\frac{3}{8} & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ -\frac{2}{3} \\ \frac{5}{8} \end{bmatrix}.$$

En segundo lugar, descomponemos el sistema en:

$$\left\{ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & -\frac{1}{2} & 0 \\ \frac{1}{6} & 0 & -\frac{1}{3} \\ \frac{1}{2} & -\frac{3}{8} & 0 \end{bmatrix} \right\} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ -\frac{2}{3} \\ \frac{5}{8} \end{bmatrix},$$

para a continuación dejar en el primer miembro el vector $(x, y, z)^t$ (producto de la identidad por $(x, y, z)^t$) y el resto en el segundo miembro

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ -\frac{1}{6} & 0 & +\frac{1}{3} \\ -\frac{1}{2} & +\frac{3}{8} & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} 1 \\ -\frac{2}{3} \\ \frac{5}{8} \end{bmatrix}$$

a partir de la cual se construye la recurrencia:

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \\ z_{k+1} \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ -\frac{1}{6} & 0 & \frac{1}{3} \\ -\frac{1}{2} & \frac{3}{8} & 0 \end{bmatrix} \begin{bmatrix} x_k \\ y_k \\ z_k \end{bmatrix} + \begin{bmatrix} 1 \\ -\frac{2}{3} \\ \frac{5}{8} \end{bmatrix}, \qquad \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix} = \begin{bmatrix} 1 \\ -\frac{2}{3} \\ \frac{5}{8} \end{bmatrix}.$$

La solución la buscaremos con tres cifras decimales exactas; así, se tiene:

n	0	1	2	3	4	5
x_n	1.000	0.667	0.688	0.590	0.619	0.613
y_n	-0.667	-0.625	-0.819	-0.762	-0.774	-0.755
z_n	0.625	-0.125	0.057	-0.026	0.044	0.025

n	6	7	8	9	10
x_n	0.622	0.620	0.620	0.620	0.620
y_n	-0.760	-0.759	-0.760	-0.760	-0.760
z_n	0.035	0.029	0.031	0.030	0.030

Por tanto, la solución es: $(x, y, z)^t = (0.620, -0.760, 0.030)$.

♦

6.3.3. Método de Gauss-Seidel

El método de Gauss-Siedel puede considerarse una variación del método de Jacobi en el sentido siguiente: en el método de Jacobi se calcula a partir de una iteración $\overrightarrow{x}^{(k)}$ la iteración $\overrightarrow{x}^{(k+1)}$, utilizándose durante todo el cálculo las componentes de la k-ésima iteración. El algoritmo mediante el que se efectúa el cálculo determina, en primer, lugar el valor de $x_1^{(k+1)}$, por lo que este valor más actualizado puede ser empleado en el cálculo de $x_2^{(k+1)}$; así, al efectuar el cálculo de $x_i^{(k+1)}$ pueden ser empleados los valores de $x_1^{(k+1)}$,...,$x_{i-1}^{(k+1)}$ en lugar de los, en principio, menos precisos $x_1^{(k)}$,...,$x_{i-1}^{(k)}$, como lo que se espera en general es una convergencia más rápida.

Formalmente, el método de Gauss-Seidel consiste en tomar como matriz de partición Q la submatriz triangular inferior de A; así, el método se plantea como

$$\begin{bmatrix} a_{1,1} & 0 & \cdots & 0 & 0 \\ a_{2,1} & a_{2,2} & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n-1,1} & a_{n-1,2} & \cdots & a_{n-1,n-1} & 0 \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n-1} & a_{n,n} \end{bmatrix} \begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ \cdots \\ x_{n-1}^{(k+1)} \\ x_n^{(k+1)} \end{bmatrix} =$$

$$= \begin{bmatrix} 0 & -a_{1,2} & \cdots & -a_{1,n-1} & -a_{1,n} \\ 0 & 0 & \cdots & -a_{2,n-1} & -a_{2,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & -a_{n-1,n} \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ \cdots \\ x_{n-1}^{(k)} \\ x_n^{(k)} \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_{n-1} \\ b_n \end{bmatrix},$$

cuya solución puede obtenerse mediante:

$$\begin{array}{lllllll} x_1^{(k+1)} = & \cdots & -\frac{a_{1,2}}{a_{1,1}}x_2^{(k)} & -\cdots & -\frac{a_{1,n-1}}{a_{1,1}}x_{n-1}^{(k)} & -\frac{a_{1,n}}{a_{1,1}}x_n^{(k)} & +\frac{b_1}{a_{1,1}} \\ x_2^{(k+1)} = & -\frac{a_{1,1}}{a_{2,2}}x_1^{(k+1)} & -\cdots & -\cdots & -\frac{a_{2,n-1}}{a_{2,2}}x_{n-1}^{(k)} & -\frac{a_{2,n}}{a_{2,2}}x_n^{(k)} & +\frac{b_2}{a_{2,2}} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_n^{(k+1)} = & -\frac{a_{n,1}}{a_{n,n}}x_1^{(k+1)} & -\frac{a_{n,2}}{a_{n,n}}x_2^{(k+1)} & -\cdots & -\frac{a_{n,n-1}}{a_{n,n}}x_{n-1}^{(k+1)} & -\cdots & +\frac{b_n}{a_{n,n}} \end{array}.$$

Ejemplo 6.3.2
Resolver mediante el método de Gauss-Seidel el sistema de ecuaciones:

$$\begin{bmatrix} 2 & -1 & 0 \\ 1 & 6 & -2 \\ 4 & -3 & 8 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2 \\ -4 \\ 5 \end{bmatrix}$$

Para resolver el problema mediante el método de Gauss-Seidel escribimos de modo explícito la iteración de Jacobi; así, se tiene:

$$\begin{aligned} x_{k+1} &= -\frac{1}{2}y_k + 1 \\ y_{k+1} &= -\frac{1}{6}x_k + \frac{1}{3}z_k - \frac{2}{3} \\ z_{k+1} &= -\frac{1}{2}x_k + \frac{3}{8}y_k + \frac{5}{8}. \end{aligned}$$

El algoritmo de Gauss-Seidel se obtiene reemplazando en el segundo miembro las variables por su nueva iteración cuando esto sea posible; así, se obtiene:

$$\begin{aligned} x_{k+1} &= -\frac{1}{2}y_k + 1 \\ y_{k+1} &= -\frac{1}{6}x_{k+1} + \frac{1}{3}z_k - \frac{2}{3} \\ z_{k+1} &= -\frac{1}{2}x_{k+1} + \frac{3}{8}y_{k+1} + \frac{5}{8} \end{aligned}$$

n	0	1	2	3	4	5	6	7	8
x_n	1.000	0.667	0.715	0.620	0.612	0.620	0.621	0.620	0.620
y_n	-0.667	-0.569	-0.760	-0.776	-0.761	-0.759	-0.760	-0.760	-0.760
z_n	0.625	0.078	-0.018	0.024	0.034	0.031	0.030	0.030	0.030

♦

6.3.4. Análisis de la convergencia de los métodos

Como es sabido, la convergencia de un método iterativo $\overrightarrow{x}^{(k+1)} = M\overrightarrow{x}^{(k)} + \overrightarrow{\beta}$ está garantizada si se verifica que existe una norma matricial subordinada a una norma vectorial tal que $||M|| \leq 1$.

Por la sencillez de su aplicación, consideramos la norma vectorial en $\mathbb{R}^n$ dada por $\|x\| = \text{máx}\{|x_1|, |x_2|, \ldots, |x_n|\}$. La norma matricial subordinada a ésta, resulta

$$\|M\| = \text{máx}\left\{\sum_{j=1}^{n} |a_{1,j}|, \sum_{j=1}^{n} |a_{2,j}|, \ldots, \sum_{j=1}^{n} |a_{n,j}|\right\}.$$

En el caso del método de Richardson la matriz de M resulta $I - A$ por lo que la condición se verifica si

$$\|M\| = \text{máx}\left\{|1 - a_{1,1}| + \sum_{j=2}^{n} |a_{1,j}|, |a_{2,1}| + |1 - a_{2,2}| + \sum_{j=3}^{n} |a_{2,j}|, \ldots, \right.$$
$$\left. \ldots, \sum_{j=1}^{n-2} |a_{n-1,j}| + |1 - a_{n-1,n-1}| + |a_{n-1,n}|, \sum_{j=1}^{n-1} |a_{n,j}| + |1 - a_{n,n}|\right\} \leq 1,$$

lo que se verifica si las componentes de la diagonal de A son suficientemente próximas a la unidad, y el resto de componentes a cero.
Respecto al método de Jacobi, la condición se cumple si se verifica

$$\sum_{j=0}^{i-1} \left|\frac{a_{i,j}}{a_{i,i}}\right| + \sum_{j=i+1}^{n} \left|\frac{a_{i,j}}{a_{i,i}}\right| \leq 1 \qquad \forall i = 1, \ldots, n,$$

lo cual es equivalente a $\sum_{j=0}^{i-1} |a_{i,j}| + \sum_{j=i+1}^{n} |a_{i,j}| \leq |a_{i,i}|, \quad \forall i = 1, \ldots, n$, lo que coincide con la definición de matriz diagonal dominante. Por tanto, una condición suficiente de convergencia para el método de Jacobi es que la matriz A sea diagonal dominante.

Respecto al método de Gauss-Seidel consideremos, en primer, lugar el sistema inicial escrito en la forma $\overrightarrow{x} = M\overrightarrow{x} + \overrightarrow{\beta}$ donde M y $\overrightarrow{\beta}$ vienen dados por

$$M = \begin{bmatrix} 0 & -\frac{a_{1,2}}{a_{1,1}} & \cdots & -\frac{a_{1,n-1}}{a_{1,1}} & -\frac{a_{1,n}}{a_{1,1}} \\ -\frac{a_{1,2}}{a_{2,2}} & 0 & \cdots & -\frac{a_{2,n-1}}{a_{2,2}} & -\frac{a_{2,n}}{a_{2,2}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -\frac{a_{n-1,1}}{a_{n-1,n-1}} & -\frac{a_{n-1,2}}{a_{n-1,n-1}} & \cdots & 0 & -\frac{a_{n-1,n}}{a_{n-1,n-1}} \\ -\frac{a_{n,1}}{a_{n,n}} & -\frac{a_{n,2}}{a_{n,n}} & \cdots & -\frac{a_{n,n-1}}{a_{n,n}} & 0 \end{bmatrix}, \quad \overrightarrow{\beta} = \begin{bmatrix} \frac{b_1}{a_{1,1}} \\ \frac{b_2}{a_{2,2}} \\ \cdots \\ \frac{b_{n-1}}{a_{n-1,n-1}} \\ \frac{b_n}{a_{n,n}} \end{bmatrix}.$$

Con esta notación el algoritmo de Gauss-Seidel puede ser escrito como

$$x_i^{(k+1)} = \sum_{j=0}^{i-1} m_{i,j} x_j^{(k+1)} + \sum_{j=i}^{n} m_{i,j} x_j^{(k)} + \beta_i, \qquad \forall i = 1, \ldots n.$$

La solución del sistema $\overrightarrow{x}$ satisface por su parte

$$x_i = \sum_{j=0}^{i-1} m_{i,j} x_j + \sum_{j=i}^{n} m_{i,j} x_j + \beta_i, \qquad \forall i = 1, \ldots n,$$

por tanto, se verifica

$$x_i - x_i^{(k+1)} = \sum_{j=0}^{i-1} m_{i,j}\left(x_i - x_j^{(k+1)}\right) + \sum_{j=i}^{n} m_{i,j}\left(x_i - x_j^{(k)}\right), \qquad \forall i = 1, \dots n,$$

y por tanto,

$$|x_i - x_i^{(k+1)}| = \sum_{j=0}^{i-1} m_{i,j}\left|x_i - x_j^{(k+1)}\right| + \sum_{j=i}^{n}\left|x_i - x_j^{(k)}\right|, \qquad \forall i = 1, \dots n.$$

Por otra parte $\left\|\overrightarrow{x} - \overrightarrow{x}^{(k)}\right\| = \text{máx}\left\{\left|x_i - x_i^{(k)}\right|; i = 1, \dots, n\right\}$ y como consecuencia se verifica $\left|x_i - x_i^{(k)}\right| \leq \left\|\overrightarrow{x} - \overrightarrow{x}^{(k)}\right\|, \quad \forall i = 1.,,,n.$

Sea $p_i = \sum\limits_{j=0}^{i-1} |a_{i,j}|$, $q_i = \sum\limits_{j=i}^{n} |a_{i,j}|$, $\forall i = 1, \dots, n$. Con esta notación se verifica

$$\left|x_i - x_i^{(k+1)}\right| \leq p_i\left\|\overrightarrow{x} - \overrightarrow{x}^{(k+1)}\right\| + q_i\left\|\overrightarrow{x} - \overrightarrow{x}^{(k)}\right\|, \quad \forall i = 1, \dots, n,$$

entonces, si s es el valor para el que es máximo el segundo miembro, también se verifica

$$\left\|\overrightarrow{x} - \overrightarrow{x}^{(k+1)}\right\| \leq p_s\left\|\overrightarrow{x} - \overrightarrow{x}^{(k+1)}\right\| + q_s\left\|\overrightarrow{x} - \overrightarrow{x}^{(k)}\right\|,$$

o lo que es equivalente,

$$\left|\overrightarrow{x} - \overrightarrow{x}^{(k+1)}\right\| \leq \frac{q_s}{1-p_s}\left\|\overrightarrow{x} - \overrightarrow{x}^{(k)}\right\|.$$

Sea $\mu = \text{máx}\left\{\frac{q_i}{1-p_i}; i = 1, \dots, n\right\}$, entonces:

$$\left\|\overrightarrow{x} - \overrightarrow{x}^{(k+1)}\right\| \leq \mu\left\|\overrightarrow{x} - \overrightarrow{x}^{(k)}\right\|.$$

Entonces se verifica que

$$\left\|\overrightarrow{x} - \overrightarrow{x}^{(k)}\right\| \leq \mu\left\|\overrightarrow{x} - \overrightarrow{x}^{(k-1)}\right\| \leq \mu^k\left\|\overrightarrow{x} - \overrightarrow{x}^{(0)}\right\|.$$

Por tanto, el proceso será convergente si $\mu < 1$. Por otra parte,

$$p_i + q_i \leq \|M\|, \quad \forall i = 1, \dots, n,$$

de donde $\forall i = 1, \dots, n$ se obtiene $q_i \leq \|M\| - p_i$. Si además se verifica $\|M\| > 1$, entonces:

$$\frac{q_i}{1-p_i} \leq \frac{\|M\| - p_i}{1-p_i} \leq \frac{\|M\| - p_i + \|M\|}{1-p_i} = \|M\|, \forall i = 1, \dots, n, \dots,$$

de donde $\mu \leq \|M\| < 1$ y, por tanto, se obtiene que si $\|M\| < 1$ el proceso de Gauss-Seidel es convergente.

Ejemplo 6.3.3
Estudiar la convergencia de los métodos de Jacobi y Gauss-Seidel para el sistema

$$\begin{bmatrix} 2 & -1 & 0 \\ 1 & 6 & -2 \\ 4 & -3 & 8 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2 \\ -4 \\ 5 \end{bmatrix}$$

Los métodos anteriores convergen para un sistema $\overrightarrow{x} = M\overrightarrow{x} + \overrightarrow{b}$ si se satisface que una de las normas matriciales (norma fila, norma columna o norma total) son menores que la unidad. Por ello, calculamos la norma fila de M definida como:

$$||M|| = \text{máx}\left\{\sum_{i=0}^{3} |a_{i,j}|; j = 1, 3\right\} = \text{máx}\left\{\frac{1}{2}, \frac{1}{6} + \frac{1}{3}, \frac{1}{2} + \frac{3}{8}\right\} =$$

$$= \text{máx}\left\{\frac{1}{2}, \frac{1}{2}, \frac{7}{8}\right\} = \frac{7}{8} < 1.$$

Por tanto, se dan las condiciones de convergencia de ambos métodos.

♦

Tema 7

Aproximación de funciones

7.1. Introducción

En el presente capítulo se pretende abordar el problema de estimar el valor de una función en un punto a partir de un conjunto de valores de la misma, en unos determinados nodos $x_0, x_1, \cdots, x_n$. Este proceso es conocido como reconstrucción de una función, y es el proceso inverso a la discretización, la cual consiste en calcular los valores de una función sobre un conjunto discreto de puntos. El proceso de discretización es único, no así el de reconstrucción pues dada la tabla de valores $\{(x_i, y_i)\}_{i=0}^n$ de una función $f(x)$, existen muchas funciones $G : \to \mathbb{R}$ de modo que $G(x_i) = y_i$, $\forall i = 0, \ldots, n$.

Existen varias técnicas para reconstruir una función, una de las más sencillas consiste en elegir una función $G(x)$ en un espacio de funciones suficientemente general de modo que $G(x_i) = y_i$, $\forall i = 0, \ldots, n$. Una función G que cumple la condición anterior se dice que interpola la tabla $\{(x_i, y_i)\}_{i=0}^n$, y cuando dicha tabla se corresponde con la de valores de una determinada función $f(x)$ se dice que $G(x)$ interpola a $f(x)$ en los nodos $x_0, x_1, \ldots, x_n$.

Los métodos de interpolación se basan en encontrar una clase de funciones que pueda aproximar tanto como se desee a una función continua en un intervalo $[a, b]$ y elegir dentro de esta clase una que cumpla ciertas condiciones e interpole la tabla. Los métodos más sencillos de interpolación son los llamados métodos de interpolación polinomial, pues los polinomios cumplen el teorema de aproximación de Weierstrass que enunciamos sin demostrar.

Teorema 7.1.1 (Teorema de Weierstrass)
Sea $f : [a, b] \to \mathbb{R}$ una función continua, entonces $\forall \epsilon \leq 0$ $\exists P(x)$ polinomio de modo que $|f(x) - P(x)| \leq \epsilon, \forall x \in [a, b]$.

Cuando los datos sobre los que se reconstruye la función no son exactos, por ejemplo, si han sido obtenidos mediante procesos de medida y pueden estar afectados de errores, es preferible recurrir a técnicas de regresión para estimar la función, dando lugar a otras técnicas de reconstrucción.

En el presente capítulo únicamente se abordara la reconstrucción mediante interpolación polinómica.

7.2. Consideraciones generales sobre la interpolación polinómica

A continuación se aborda el estudio de una serie de consideraciones generales sobre la interpolación polinómica.

Teorema 7.2.1
Sea una tabla de valores $\{(x_i, y_i)\}_{i=0}^{n}$ con $x_0 < x_1 < \ldots < x_n$. Entonces, existe un único polinomio de grado menor o igual que n que interpola la tabla.

Dem:
Un polinomio de grado menor o igual que n puede representarse como $P_n(x) = \sum_{i=0}^{n} a_i x^i$. Por tanto, el polinomio queda unívocamente determinado cuando se conocen sus coeficientes. El polinomio interpola la tabla en los nodos $x_0, x_1, \ldots, x_n$, por tanto, se satisfacen las ecuaciones

$$\begin{aligned} P_n(x_0) &= a_0 + a_1 x_0 + a_2 x_0^2 + \cdots + a_n x_0^n = y_0 \\ P_n(x_1) &= a_0 + a_1 x_1 + a_2 x_1^2 + \cdots + a_n x_1^n = y_1 \\ P_n(x_2) &= a_0 + a_1 x_2 + a_2 x_2^2 + \cdots + a_n x_2^2 = y_2 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ P_n(x_n) &= a_0 + a_1 x_n + a_2 x_n^2 + \cdots + a_n x_0^n = y_0 \end{aligned}$$

lo que escrito en forma matricial resulta

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

El sistema es compatible y determinado, puesto que su determinante resulta el conocido determinante de Vandermonde cuyo valor viene dado por

$$\begin{vmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{vmatrix} = \prod_{i=1}^{n} \prod_{j=0}^{i-1} (x_i - x_j) \neq 0$$

pues $x_0 < x_1, \cdots < x_n$. Por tanto, se verifica el teorema.

▼

En general, dada una función $f(x)$ de la cual disponemos de la tabla de valores $\{(x_i, y_i)\}_{i=0}^{n}$ sobre los nodos $x_0 < x_1 < \cdots < x_n$, y dado un polinomio que interpola dicha tabla, únicamente podemos afirmar que $f(x)$ y $P(x)$ coinciden en los nodos. Para un punto x distinto a los nodos, en general podemos decir que el error $E(x) = f(x) - P_n(x)$ cometido al aproximar $f(x)$ mediante $P_n(x)$ será habitualmente $E(x) \neq 0$. La función $E(x)$ se puede acotar en algunos casos si se tiene más

información sobre la función $f(x)$. El siguiente teorema establece una acotación del error cuando la función a interpolar es suficientemente derivable.

Teorema 7.2.2
Sea $f : [a, b] \to \mathbb{R}$, $f \in C^{n+1}[a, b]$, y sean los nodos $a \leq x_0 < x_1 < \ldots. < x_n \leq b$. Entonces se verifica que la función $E_n(x) = f(x) - P_n(x)$ donde $P_n(x)$ es el polinomio interpolante de grado menor o igual que n satisface la acotación.

$$|E_n(x)| = \frac{\text{máx}\left\{\left|f^{n+1)}(x)\right|; x \in [a, b]\right\}}{(n+1)!}|x - x_0||x - x_1| \cdot \cdots \cdot |x - x_n|.$$

Dem:

Sea $f \in C^{n+1}([a, b])$ y sea $P_n(x)$ el polinomio que interpola dicha función sobre los nodos $a \leq x_0 < x_1 < \cdots < x_{n-1} < x_n \leq b$.

Sea $x \in [x_0, x_n]$ y sea $E(x) = f(x) - P_n(x)$ el error cometido cuando se aproxima $f(x)$ mediante $P_n(x)$. Si x coincide con alguno de los nodos, dicho error es cero. Por otra parte, sea $\lambda \in \mathbb{R}$ el valor para el cual una vez fijado x se tiene

$$E(x) = \lambda(x - x_0)(x - x_1)\ldots(x - x_n).$$

Sea $x_0 < x_1 < \cdots < x_{k-1} < x < x_k < \cdots < x_n$, y sea $\Psi(x)$ la función definida como

$$\Psi(t) = f(t) - P_n(t) - \lambda(t - x_0)(t - x_1)\cdots(t - x_n).$$

Dicha función es de clase $C^{k+1}([a, b])$ y satisface que $\Psi(x_0) = \Psi(x_1) = 0$. Por tanto, por el teorema de Rolle existe $\xi_{0,0} \in]x_0, x_1[$ de modo que $\Psi'(\xi_{0,0}) = 0$. Análogamente, existe $\xi_{0,1} \in]x_1, x_2[$ de modo que $\Psi'(\xi_{0,1}) = 0, \ldots, \xi_{0,k-1} \in]x_{k-1}, x[$ de modo que $\Psi'(\xi_{0,k-1}) = 0$, $\xi_{0,k} \in]x, x_k[$ de modo que $\Psi'(\xi_{0,k}) = 0, \ldots, \xi_{0,n} \in]x_{n-1}, x_n[$, $\Psi'(\xi_{0,n}) = 0$. Dichos puntos satisfacen $a < \xi_{0,0} < \xi_{0,1} < \cdots < \xi_{0,n} < b$. En el intervalo $[\xi_{0,i}, \xi_{0,i+1}]$ se tiene que $\Psi'(\xi_{0,i}) = \Psi'(\xi_{0,1+1}) = 0$ y $\Psi' \in C^k([\xi_{0,i}, \xi_{0,i+1}])$ por tanto, $\exists \xi_{1,i} \in]\xi_{0,i}, \xi_{0,i+1}[$ de modo que $\Psi''(\xi_{1,i}) = 0, \forall i = 0, \ldots, n-1$.

Repitiendo el proceso se obtiene que $\exists \xi_{n,0} \in]\xi_{n-1,0}, \xi_{n-1,1}[$ de manera que se verifica $\left.\frac{d^{n+1}\Psi(t)}{dt^{n+1}}\right|_{t=\xi_{n,0}} = 0$, por tanto $\exists \xi \in]a, b[$ de modo que $\Psi^{n+1)}(\xi) = 0$, en particular $\xi = \xi_{n,0}$.

Por otra parte, se tiene que $\Psi^{n+1)}(t) = f^{n+1)}(t) - \lambda(n+1)!$ de donde se tiene que $\lambda = \frac{f^{n+1)}(xi)}{(n+1)!}$, y por tanto, $\exists \xi \in]a, b[$ de modo que

$$E(x) = \frac{f^{n+1)}(\xi)}{(n+1)!}(x - x_0)(x - x_1)\cdots(x - x_n),$$

a partir de lo cual se obtiene

$$|E(x)| = \frac{|f^{n+1)}(\xi)|}{(n+1)!}|x - x_0| \cdot |x - x_1| \cdots |x - x_n|,$$

y por tanto:

$$|E(x)| \leq \frac{\text{máx}\{|f^{n+1)}(\xi)| : \xi \in [a, b]\}}{(n+1)!}|x - x_0| \cdot |x - x_1| \cdots |x - x_n|.$$

▼

Ejemplo 7.2.1
Determinar el máximo error que se comete al aproximar el valor de $\operatorname{sen} 0.4$ a partir de los valores de la función seno en los nodos $0, 0.25, 0.5, 0.75, 1$

En este caso, se tiene que $n = 5$, y $\frac{d^5 \operatorname{sen} x}{dx^5} = \cos x$; así, se tiene

$$|E(x)| \leq \frac{1}{5!}|0.4 - 0.0| \cdot |0.4 - 0.25| \cdot |0.4 - 0.5| \cdot |0.4 - 0.75| \cdot |0.4 - 1.0| =$$

$$= \frac{1}{120} 0.4 \cdot 0.14 \cdot 0.1 \cdot 0.36 \cdot 0.6 = 0.0000105.$$

♦

La expresión del error consta de dos partes, la primera $\frac{|\max\{|f^{n+1)}(\xi)|:\xi\in[a,b]\}}{(n+1)!}$ depende del número de nodos y de la función a interpolar. La segunda parte, $|x - x_0| \cdots |x - x_1| \cdots |x - x_n|$ depende de los nodos y del punto en el que se interpola.

Para tratar de minimizar el error en la interpolación de orden n, nada se puede hacer respecto a la primera parte del error, pues la función cuyos valores se interpolan viene dada. Respecto a la segunda parte, y si es posible elegir los nodos sobre los que se realiza la interpolación, estos deben ser elegidos de modo que se minimice el máximo valor que puede tomar $|x - x_0| \cdots |x - x_1| \cdots |x - x_n|$ para $x \in [a, b]$. Para ello se introducen a continuación los llamados polinomios de Tchevichev.

Definición 7.2.1
Llamamos **polinomio de Tchevichev** de grado n a $T_n(x) = \cos(n \arccos x)$.

Teorema 7.2.3
Las funciones $T_n(x)$ son polinomios cuyo valor viene dado por $T_0(x) = 1$, $T_1(x) = x$, $T_{n+2}(x) = 2xT_{n+1}(x) - T_n(x)$, $n = 0, 1, \ldots$

Dem:

Para $n = 0$ se tiene $T_0(x) = \cos(0 \arccos x)) = \cos 0 = 1$. Para $n = 1$ se tiene $T_1(x) = \cos(\arccos x) = x$.

$$T_{n+2}(x) = \cos[(n + 2)\arccos x] = \cos[(n + 1)\arccos x + \arccos x] =$$
$$= \cos[(n + 1)\arccos x] \cos[\arccos x] - \operatorname{sen}[(n + 1)\arccos x]\operatorname{sen}[\arccos x]$$

$$T_{n+2}(x) = \cos[(n + 2)\arccos x] = T_{n+1}(x)\, x - \operatorname{sen}[(n + 1)\arccos x]\operatorname{sen}[\arccos x].$$

Por otra parte, se tiene que

$$\operatorname{sen} x \operatorname{sen} y = -\frac{1}{2}\cos(x + y) + \frac{1}{2}\cos(x - y)$$

y por tanto,

$$\operatorname{sen}[(n+1)\arccos x]\operatorname{sen}[\arccos x] = -\frac{1}{2}\cos[(n+2)\arccos x] + \frac{1}{2}\cos[n \arccos x] =$$
$$= -\frac{1}{2}T_{n+2}(x) + \frac{1}{2}T_n(x).$$

Así:

$$T_{n+2}(x) = \cos[(n+2)\arccos x] = xT_{n+1}(x) + \frac{1}{2}T_{n+2}(x) - \frac{1}{2}T_n(x)$$

de donde

$$T_{n+2}(x) = 2xT_{n+1}(x) - T_n(x),\ n \geq 0$$

▼

Definición 7.2.2
Sea $f : [a,b] \to \mathbb{R}$. Llamamos **norma infinito** de f a $||f||_\infty = \text{máx}\{|f(x)| : x \in [a,b]\}$.

Teorema 7.2.4
Sea $P_n : [-1,1] \to \mathbb{R}$ una función polinómica de grado n cuyo coeficiente principal es igual a la unidad. Entonces $||P_n||_\infty \geq 2^{1-n}$.

Dem:

Supongamos que existe una función polinómica $P_n(x) = a_n x^n + \ldots + a_1 x + a_0$ con $a_n = 1$, cuyo dominio está restringido a $[-1,1]$ de modo que $||P_n||_\infty < 2^{1-n}$. Sea $Q(x) = 2^{1-n}T_n(x)$, entonces $Q(x)$ es un polinomio de grado n cuyo coeficiente principal es la unidad, pues según la fórmula de recurrencia obtenida en el teorema anterior, el coeficiente principal de $T_n(x)$ es 2^{n-1}.

Sean los valores $x_k = \cos\frac{k\pi}{n}$ $k = 0, \ldots, n$, los cuales están comprendidos en $[-1,1]$. Por otra parte:

$$Q(x_k) = 2^{1-n}\cos\left[n\text{arccos}\left(\cos\frac{k\pi}{n}\right)\right] = 2^{1-n}\cos k\pi = (-1)^k 2^{1-n}$$

Sea el polinomio $S(x) = Q(x) - P_n(x)$ entonces se verifica que $S(x)$ es un polinomio de grado menor o igual que n. También se verifica

$$(-1)^k P_n(x_k) \leq |P_n(x_k)| < 2^{1-n} = (-1)^k T_n(x_k),$$

de donde

$$(-1)^k S(x_k) = (-1)^k[Q(x) - P_n(x)] > 0,$$

por tanto, $S(x)$ presenta al menos n cambios de signo en el intervalo $[-1,1]$ y por continuidad, n raíces, lo cual no es posible pues el grado de $S(x)$ es menor o igual que $n-1$, por tanto, no es posible admitir $||P_n||_\infty < 2^{1-n}$, de donde se sigue el teorema. ▼

7.3. Interpolación de Lagrange

Tal y como se ha visto anteriormente, dada una tabla $\{(x_i, y_i)\}_{i=0}^n$ de valores de una función sobre los nodos $x_0 < x_1 < \cdots < x_n$, existe un único polinomio de grado menor o igual que n que interpola a la función en dichos nodos. En el teorema de existencia y unicidad se proporciona un método para la obtención del

polinomio interpolante; este método consiste en la resolución del sistema de ecuaciones lineales que satisfacen sus coeficientes. Sin embargo, este procedimiento no es práctico, pues la resolución del sistema implica un gran número de operaciones. Para solventar este problema se proponen varios métodos para obtener el polinomio interpolante sin recurrir a la resolución de dicho sistema, entre tales métodos están los métodos de Lagrange, de Newton, etc. Nótese que todos los métodos conducen al mismo polinomio interpolante, y son únicamente, distintos procedimientos para su obtención.

En primer lugar, abordamos el estudio del método de Lagrange para la obtención del polinomio de interpolación. Este método está basado en el siguiente teorema:

Teorema 7.3.1
Sea $x_0 < x_1 < \cdots < x_n$ un conjunto de nodos. Entonces, existe un único conjunto de funciones $\{\ell_i(x)\}_{i=0}^n$ tales que satisfacen:

1.- $\{\ell_i(x)\}_{i=0}^n$ son polinomios de grado n.
2.- $\ell_i(x_i) = 1, \forall i = 0, 1, \ldots, n.$
3.- $\ell_i(x_j) = 0, \forall i, j = 0, 1, \ldots, n; \quad i \neq j.$

Dem:

La función $\ell_i(x_j) = 0, \forall i, j = 0, 1, \ldots, n$ si $i \neq j$, por tanto, x_j es raíz de $\ell_i(x)$ cuando $j \neq i$; así, $\ell_i(x) = \phi_i(x)(x - x_0) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)$.
Puesto que $\ell_i(x)$ debe ser un polinomio de grado n, $\phi_i(x) = C_i \in \mathbb{R}$ debe ser una constante real, y por tanto $\ell_i(x) = C_i(x - x_0) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)$.
El valor de la constante lo determinamos a partir de la condición $\ell_i(x_i) = 1$. Así:

$$C_i(x_i - x_0) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n) = 1,$$

y por tanto

$$C_i = \frac{1}{(x_i - x_0) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)},$$

de donde

$$\ell_i(x) = \frac{(x - x_0) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)}.$$

Nótese que dichas funciones dependen únicamente de los nodos y no de los valores de la función a interpolar.
El conjunto de funciones $\{\ell_i(x)\}_{i=0}^n$ recibe el nombre de funciones de base de la interpolación de Lagrange. Estas funciones constituyen una base del espacio vectorial de los polinomios de grado menor o igual que n, pues constituyen un conjunto linealmente independiente de polinomios de grado menor o igual que n de $n + 1$ elementos (igual a la dimensión del espacio vectorial que se pretende generar) y, por tanto, constituyen una base. La independencia lineal se desprende inmediatamente, pues sea $a_0, a_1, \ldots, a_n \in \mathbb{R}$ de modo que se satisfaga $\sum_{i=0}^{n} a_i \ell_i(x) = 0$,

así $\sum_{i=0}^{n} a_i\ell_i(x_j) = 0$, $\forall j = 0, 1, \ldots, n$ y por tanto:

$$\sum_{i=0}^{n} a_i\ell_i(x_j) = a_0\ell_0(x_j) + \ldots + a_{j-1}\ell_{j-1}(x_j) +$$
$$+ a_j\ell_j(x_j) + a_{j+1}\ell_{j+1}(x_j) + \ldots + a_n\ell_n(x_j) =$$
$$= a_0 0 + \ldots + a_{j-1}0 + a_j 1 + a_{j+1}0 + \ldots + a_n 0 = a_j = 0.$$

Entonces, $a_j = 0, \quad \forall j = 0, 1, \ldots, n.$
El polinomio de interpolación puede ser construido fácilmente a partir de dichas funciones como

$$L_n(x) = \sum_{i=0}^{n} y_i\ell_i(x).$$

Dicho polinomio recibe el nombre de polinomio interpolante de Lagrange. Veamos que efectivamente dicho polinomio interpola la tabla $\{(x_i, y_i)\}_{i=0}^{n}$, para ello sea un nodo arbitrario $x_j \in \{x_0, x_1, \ldots, x_n\}$, entonces:

$$L_n(x_j) = \sum_{i=0}^{n} y_i\ell_i(x_j) = y_0\ell_0(x_j) + \ldots + y_{j-1}\ell_{j-1}(x_j) +$$
$$+ y_j\ell_j(x_j) + y_{j+1}\ell_{j+1}(x_j) + \ldots + yn\ell_n(x_j) =$$
$$= ay_0 + \ldots + y_{j-1}0 + y_j 1 + y_{j+1}0 + \ldots + y_n 0 = y_j,$$

y por tanto, dicho polinomio interpola la tabla.

▼

Finalmente, veamos una notación más compacta para dichos polinomios. Definamos en primer lugar la función $\Pi(x) = \prod_{i=0}^{n}(x - x_i)$ y calculemos su derivada la cual podemos expresar como:

$$\Pi'(x) = \begin{cases} \sum\limits_{i=0} \frac{\Pi(x)}{x-x_i} & x \notin \{x_0, x_1, \ldots, x_n\} \\ (x_i - x_0)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_n) & x = x_i, \quad \forall i = 0, 1, \ldots, n. \end{cases}$$

con lo cual se tiene que $L_n(x) = \sum i = 0^n \frac{\Pi(x)}{(x-x_i)\Pi'(x_i)} y_i$ si $x \notin \{x_0, x_1, \ldots, x_n\}$, y por tanto:

$$L_n(x) = \begin{cases} \sum\limits_{i=0}^{n} \frac{\Pi(x)}{(x-x_i)\Pi'(x_i)} y_i & x \notin \{x_0, x_1, \ldots, x_n\} \\ y_i & x = x_i, \quad i = 0, 1, \ldots n \end{cases}$$

Ejemplo 7.3.1
Dada la tabla

n	0	1	2	3
x_n	1	2	0	3
y_n	3	2	-4	5

Calcular mediante el método de interpolación de Lagrange los valores $f(1.5)$ y $f(2.5)$.

Para calcular el polinomio de interpolación de Lagrange calcularemos en cada caso los valores de la tabla:

$$\begin{array}{cccc} x-x_0 & x_0-x_1 & x_0-x_2 & x_0-x_3 \\ x_1-x_0 & x-x_1 & x_1-x_2 & x_1-x_3 \\ x_2-x_0 & x_2-x_1 & x-x_2 & x_2-x_3 \\ x_3-x_0 & x_3-x_1 & x_3-x_2 & x-x_3, \end{array}$$

lo que resulta para $x = 1.5$:

$$\begin{array}{ccccccccc} 0.5 & -1 & 1 & -2 & \rightarrow & 1 & \rightarrow & l_0(1.5) = \frac{0.5625}{1} = 0.5625 \\ 1 & -0.5 & 2 & -1 & \rightarrow & 1 & \rightarrow & l_1(1.5) = \frac{0.5625}{1} = 0.5625 \\ -1 & -2 & 1.5 & -3 & \rightarrow & -9 & \rightarrow & l_2(1.5) = \frac{0.5625}{-9} = -0.0625 \\ 2 & 1 & 3 & -1.5 & \rightarrow & -9 & \rightarrow & l_3(1.5) = \frac{0.5625}{-9} = -0.0625 \\ & & & & \searrow & & & \\ & & & & & 0.5625 & & \end{array}$$

En la tabla anterior detrás de las flechas aparecen el producto de las filas y de la diagonal, respectivamente. En la última columna, aparece el valor de las funciones de base en el punto $x = 1.5$.

Finalmente, el valor del polinomio de Lagrange lo calcularemos como:

$$L_3(1.5) = \sum_{i=0}^{3} l_i(1.5)y_i = 0.5625\cdot 3 + 0.5625\cdot 2 - 0.0625\cdot(-4) - 0.0625\cdot 5 = 2.75\,.$$

Para interpolar en el punto $x = 2.5$ por el método de Lagrange bastará cambiar la diagonal de la tabla y efectuar los correspondientes cálculos:

$$\begin{array}{ccccccccc} 1.5 & -1 & 1 & -2 & \rightarrow & 3 & \rightarrow & l_0(2.5) = \frac{-0.9375}{3} = -0.3125 \\ 1 & 0.5 & 2 & -1 & \rightarrow & -1 & \rightarrow & l_1(2.5) = \frac{-0.9375}{-1} = 0.9375 \\ -1 & -2 & 2.5 & -3 & \rightarrow & -15 & \rightarrow & l_2(2.5) = \frac{-0.9375}{-15} = 0.0625 \\ 2 & 1 & 3 & -0.5 & \rightarrow & -3 & \rightarrow & l_3(2.5) = \frac{-0.9375}{-3} = 0.3125 \\ & & & & \searrow & & & \\ & & & & & -0.9375 & & \end{array}$$

$$L_3(2.5) = \sum_{i=0}^{3} l_i(2.5)y_i = -0.3125\cdot 3 + 0.9375\cdot 2 + 0.0625\cdot(-4) + .3125\cdot 5 = 2.25\,.$$

♦

7.4. Interpolación de Newton

En esta sección se aborda el estudio de un método alternativo al de Lagrange para la obtención de un polinomio interpolante de grado menor o igual que n para la tabla de valores $\{(x_i, y_i)\}_{i=0}^{n}$. Para ello definamos, en primer lugar, el concepto de diferencia dividida de una función $f(x)$ sobre los puntos x_i, x_j.

Definición 7.4.1
Llamamos **diferencia dividida** de la función $f(x)$ en los puntos x_i, x_j a la cantidad

$$[x_i, x_j]_f = \frac{f(x_j) - f(x_i)}{x_j - x_i}.$$

Cuando todos los cálculos se refieren a una misma función, se suele suprimir el subíndice f, pues en ese caso no hay ambigüedad. En este apartado, puesto que se trata de interpolar una tabla de valores de una determinada función, prescindiremos del subíndice f.
Las diferencias divididas satisfacen la propiedad simétrica. Así, $[x_i, x_j] = [x_j, x_i]$.
A continuación extendemos el concepto de diferencia dividida como:

1. Diferencia dividida de orden cero

$$[x_i] = f(x_i)$$

2. Diferencia dividida de orden $k+1$

$$\left[x_{i_1}, x_{i_2}, \ldots, x_{i_k}, x_{i_{k+1}}\right] = \frac{\left[x_{i_2}, \ldots, x_{i_k}, x_{i_{k+1}}\right] - \left[x_{i_1}, x_{i_2}, \ldots, x_{i_k}\right]}{x_{i_{k+1}} - x_{i_1}}$$

A partir de esta definición se obtiene inmediatamente el siguiente resultado:
Sean los puntos $x_1, x_2, \ldots, x_n \in \mathbb{R}$, y sea σ una permutación de n elementos. Entonces $[x_1, x_2, \ldots, x_n] = \left[x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(n)}\right]$.

Teorema 7.4.1
Sea $P(x)$ un polinomio de grado n, entonces sus diferencias divididas de orden n son constantes.

Dem:
Sea $x_{i_1} \in \mathbb{R}$, y sea $[x, x_{i_1}] = \frac{P_n(x) - P_n(x_{i_1})}{x - x_{i_1}}$. Evidentemente, se tiene que x_{i_1} es raíz de $P_n(x) - P_n(x_{i_1})$, por tanto, $P_n(x) - P_n(x_{i_1})$ es divisible por $x - x_{i_1}$. Entonces la diferencia dividida $[x, x_{i_1}]$ es un polinomio de grado $n-1$.
Sea $x_{i_2} \in \mathbb{R}\, \{x_{i_1}\}$ y sea $P_{n-1}(x) = [x, x_{i_1}]$; por tanto, se satisface la relación

$$P_n(x) = P_n(x_{i_1}) + P_{n-1}(x)(x - x_{i_1}).$$

Sea la diferencia $[x, x_{i_1}, x_{i_2}]$, entonces se verifica que el numerador de

$$[x, x_{i_1}, x_{i_2}] = \frac{[x,x_{i_1}] - [x_{i_1}, x_{i_2}]}{x - x_{i_2}},$$

se anula si $x = x_{i_2}$ lo que implica que es múltiplo de $x - x_{i_2}$, por tanto, $[x, x_{i_1}, x_{i_2}]$ es un polinomio de grado $n-2$ que llamaremos $P_{n-2}(x)$. Dicho polinomio satisface la relación

$$P_{n-1}(x) = P_{n-1}(x_{i_2}) + P_{n-2}(x)(x - x_{i_2}) = [x_{i_1} - x_{i_2}] + P_{n-2}(x)(x - x_{i_2}).$$

Supongamos que para orden k se tienen los puntos distintos $x_{i_1}, \ldots, x_{i_k} \in \mathbb{R}$, y que $P_{n-k}(x) = [x, x_{i_1}, \ldots, x_{i_k}]$ es un polinomio de grado $n-k$ que satisface

$$P_{n-k}(x) = P_{n-k}(x_{i_k}) + P_{n-k+1}(x)(x - x_{i_k}) = [x_{i_1}, \ldots, x_{i_k}] + P_{n-k+1}(x - x_{i_k}).$$

Sea $x_{k+1} \in \mathbb{R} - \{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\}$. Entonces:

$$P_{n-k-1}(x) = \left[x, x_{i_1}, \ldots, x_{i_k}, x_{i_{k+1}}\right] = \frac{[x, x_{i_1}, \ldots, x_{i_k}] - \left[x_{i_1}, \ldots, x_{i_k}, x_{i_{k+1}}\right]}{x - x_{i_{k+1}}},$$

de donde

$$P_{n-k-1}(x) = \frac{P_{n-k}(x) - P_{n-k}(x_{i_{k+1}})}{x - x_{i_{k+1}}}$$

se anula en $x = x_{i_{k+1}}$ y, por tanto, $P_{n-k}(x) - P_{n-k}(x_{i_{k+1}})$ es divisible por $x - x_{i_{k+1}}$. Así, $P_{n-k-1}(x)$ es un polinomio de grado $n-k-1$ que satisface la relación

$$P_{n-k}(x) = P_{n-k-1}(x_{i_{k+1}}) + P_{n-k-1}(x)(x - x_{i_{k+1}}).$$

Por tanto, $P_0(x) = [x, x_{i_1}, x_{i_2}, \ldots, x_{i_n}] = \frac{P_1(x) - P_1(x_{i_n})}{x - x_{i_n}}$ es un polinomio de grado 0, y por consiguiente, constante. Sea $x_{i_{n+1}} \in \mathbb{R} - \{x_{i_1}, x_{i_2}, \ldots, x_{i_n}\}$. Entonces puesto que $P_0(x)$ es constante se satisface que

$$[x, x_{i_1}, x_{i_2}, \ldots, x_{i_n}] = P_0(x) = \left[x_{i_1}, x_{i_2}, \ldots, x_{i_n}, x_{i_{n+1}}\right].$$

▼

Teorema 7.4.2
El polinomio $N_n(x)$ definido como

$$N_n(x) = [x_0] + [x_0, x_1](x - x_0) + [x_0, x_1, x_2](x - x_0)(x - x_1) + \ldots + \\ + \ldots + [x_0, x_1, \ldots, x_n](x - x_0)(x - x_1) \cdots (x - x_{n-1})$$

interpola la tabla $\{(x_i, y_i)\}_{i=0}^{n}$ de valores de una función sobre los nodos $x_0 < x_1 < \cdots < x_n$.

Dem:

Aplicando el teorema anterior al polinomio interpolante de Lagrange $P_n(x) = L_n(x)$ y a los puntos $(x_{i_1}, x_{i_2}, \ldots, x_{i_{n+1}}) = (x_0, x_1, \ldots, x_n)$ se tiene que:

$$P_1(x) = P_1(x_n) + P_0(x_n)(x - x_{n-1}) = [x_0, x_1, \ldots, x_{n-1}] + [x_0, x_1, \ldots, x_n](x - x_{n-1})$$

$$P_2(x) = P_2(x_n - 1) + P_1(x)(x - x_{n-2}) = \\ = [x_0, x_1, \ldots, x_{n-2}] + \left([x_0, x_1, \ldots, x_{n-1}] + [x_0, x_1, \ldots, x_n](x - x_{n-1})\right)(x - x_{n-2}).$$

Así:

$$P_n(x) = P_n(x_0) + P_{n-1}(x)(x - x_0) = [x_0] + [x_0, x_1](x - x_0) + \\ + [x_0, x_1, x_2](x - x_0)(x - x_1) + \ldots + [x_0, x_1, \ldots, x_n](x - x_0)(x - x_1) \cdots (x - x_{n-1}).$$

Esta forma del polinomio interpolante se conoce como primera fórmula de Newton o fórmula de interpolación de Newton hacia adelante.

▼

Ejemplo 7.4.1
Dada la tabla

n	0	1	2	3
x_n	1	2	0	3
y_n	3	2	-4	5

Calcular los valores $f(1.5)$ y $f(2.5)$ mediante el polinomio de interpolación de Newton en su primera forma.

Para interpolar mediante el método de Newton calcularemos las diferencias divididas;

$$\begin{array}{c|lllll}
x_0 & [x_0] & & & \\
 & & [x_0,x_1] & & \\
x_1 & [x_1] & & [x_0,x_1,x_2] & \\
 & & [x_1,x_2] & & [x_0,x_1,x_2,x_3] \\
x_2 & [x_2] & & [x_1,x_2,x_3] & \\
 & & [x_2,x_3] & & \\
x_3 & [x_3] & & &
\end{array}$$

lo que para nuestros datos resulta:

$$\begin{array}{c|cccc}
1 & 3 & & & \\
 & & -1 & & \\
2 & 2 & & -4 & \\
 & & 3 & & 2 \\
0 & -4 & & 0 & \\
 & & 3 & & \\
3 & 5 & & &
\end{array}$$

El polinomio de interpolación de Newton (primera forma) se escribe como:

$$\begin{aligned}
N_3(x) = [x_0] + [x_0,x_1](x-x_0) + [x_0,x_1,x_2](x-x_0)(x-x_1) + \\
+ [x_0,x_1,x_2,x_3](x-x_0)(x-x_1)(x-x_2),
\end{aligned}$$

lo que resulta para $x = 1.5$:

$$N_3(1.5) = 3 - 1 \cdots 0.5 - 4 \cdots 0.5 \cdots (-0.5) + 2 \cdots 0.5 \cdots (-0.5) \cdots 1.5 = 2.75,$$

y para $x = 2.5$:

$$N_3(2.5) = 3 - 1 \cdots 1.5 - 4 \cdots 1.5 \cdots (0.5) + 2 \cdots 1.5 \cdots (0.5) \cdots 2.5 = 2.25\,.$$

♦

Teorema 7.4.3
El polinomio $N_n(x)$ dado por

$$\begin{aligned}
N_n(x) = [x_n] + [x_{n-1},x_n](x-x_n) + [x_{n-2},x_{n-1},x_n](x-x_{n-1})(x-x_n) + \ldots + \\
+ \ldots + [x_0,x_1,\ldots,x_n](x-x_1)\cdots(x-x_{n-1})(x-x_n)
\end{aligned}$$

interpola la tabla $\{(x_i,y_i)\}_{i=0}^{n}$ de valores de una función sobre los nodos $x_0 < x_1 < \cdots < x_n$.

Dem:

Se obtiene igual que en el caso anterior sin más que escribir los nodos en el orden $x_n, x_{n-1}, \ldots, x_1, x_0$.

▼

Esta fórmula es conocida como segunda forma del polinomio interpolante de Newton o también como fórmula de interpolación de Newton hacia atrás.

7.4.1. Interpolación con puntos igualmente espaciados

Definición 7.4.2
Sean los nodos $x_0 < x_1 < \cdots < x_n$ igualmente espaciados mediante un paso h, esto es, $x_i - x_{i-1} = h > 0$, $i = 1, \ldots, n$, y sean $y_0, \ldots, y_n$ los valores de una determinada función sobre ellos. Llamamos **diferencia avanzada** de la función en el nodo i a $\Delta y_i = y_{i+1} - y_i$ donde $i = 0, \ldots, n-1$.

Definición 7.4.3
Dada la tabla de valores $\{x_i, y_i\}_{i=0}^n$ de una función sobre los nodos igualmente espaciados $x_0 < x_1 < \cdots < x_n$ se define **diferencia avanzada de orden** k como $\Delta^k y_i = \delta^{k-1} y_{i+1} - \Delta^{k-1} y_i$ siendo $\Delta^0 y_i = y_i$.

Es inmediato que $\Delta^1 y_i = \Delta y_i$.

De modo análogo, se definen diferencias retardadas como:

Definición 7.4.4
Dada la tabla de valores $\{x_i, y_i\}_{i=0}^n$ de una función sobre los nodos igualmente espaciados $x_0 < x_1 < \cdots < x_n$ se define **diferencia retardada de orden** k como $\nabla^k y_i = \nabla^{k-1} y_{i+1} - \nabla^{k-1} y_i$ siendo $\nabla^0 y_i = y_i$.

A las diferencias $\nabla^1 y_i$ también se les denota como ∇y_i y normalmente son denominadas de un modo más simple como **diferencia retardada** de y en el nodo i.

Definición 7.4.5
Sea un conjunto de nodos igualmente espaciados $x_0 < x_1 < \cdots < x_n$. Llamamos **productos generalizados de orden** $k = 0, \ldots, n$ **hacia adelante y hacia atrás** a las cantidades $[x]_k$ y $\{x\}_k$ definidas como

$$[x]_k = [x]_{k-1}(x-k+1), \ \{x\}_k = \{x\}_{k-1}(x-n+k-1), \quad \text{donde } [x]_0 = \{x\}_0 = 1\,.$$

Teorema 7.4.4
Dada la tabla de valores $\{x_i, y_i\}_{i=0}^n$ de una función sobre los nodos igualmente espaciados $x_0 < x_1 < \cdots < x_n$, se tiene que el polinomio interpolante de Newton $N_n(x)$ en su primera forma viene dado por

$$N_n(x) = \sum_{k=0}^{n} \frac{\Delta^k y_0}{k!}[q]_k,$$

donde $q = \frac{x-x_0}{h}$.

Dem:

Sea $N_n(x)$ el polinomio de interpolación de Newton en su primera forma, el cual viene dado por

$$N_n(x) = [x_0] + [x_0, x_1](x - x_0) + [x_0, x_1, x_2](x - x_0)(x - x_1) + \cdots +$$
$$+ \cdots + [x_0, \ldots, x_n](x - x_0) \cdots (x - x_{n-1})$$

Según las definiciones anteriores, se tiene que $[x_i] = \Delta_i^0$, $[x_i, x_{i+1}] = \frac{y_{i+1}-y_i}{x_{i+1}-x_i} = \frac{\Delta^1 y_i}{h}$. Para evaluar las diferencias segundas, $[x_i, x_{i+1}, x_{i+2}]$, se tiene:

$$[x_i, x_{i+1}, x_{i+2}] = \frac{[x_{i+1}, x_{i+2}] - [x_i, x_{i+1}]}{x_{i+2} - x_i} = \frac{\Delta^1 y_{i+1} - \Delta^1 y_i}{2h^2} = \frac{\Delta^2 y_i}{2h^2}.$$

Supongamos que $[x_i, x_{i+1}, \ldots, x_{i+k}] = \frac{\Delta^k y_i}{k!h^k}$, entonces:

$$[x_i, x_{i+1}, \ldots, x_{i+k}, x_{i+k+1}] = \frac{[x_{i+1}, \ldots, x_{i+k}, x_{i+k+1}] - [x_i, x_{i+1}, \ldots, x_{i+k}]}{x_{i+k+1} - x_i} =$$

$$= \frac{\frac{\Delta^k y_{i+1}}{k!h^k} - \frac{\Delta^k y_i}{k!h^k}}{(k+1)h} = \frac{\Delta^{k+1} y_i}{(k+1)!h^{k+1}},$$

lo que confirma la hipótesis de inducción.

Por otra parte, se tiene que

$$\prod_{i=0}^{k}(x - x_k) = \prod_{i=0}(qh + x_0 - x_k) = \prod_{i=0}(qh - kh) = [q]_k h^k$$

por tanto:

$$N_n(x) = \sum_{k=0}^{n} \frac{\Delta^k y_0}{k!h^h}[q]_k h^k = \sum_{k=0}^{n} \frac{\Delta^k y_0}{k!}[q]_k$$

con lo que se tiene el resultado. ▼

Teorema 7.4.5

Dada la tabla de valores $\{x_i, y_i\}_{i=0}^n$ de una función sobre los nodos igualmente espaciados $x_0, \ldots, x_n$, se tiene que el polinomio interpolante de Newton $N_n(x)$ en su segunda forma viene dado por

$$N_n(x) = \sum_{k=0}^{n} \frac{\nabla^k y_n}{k!}\{q\}_k,$$

donde $q = \frac{x-x_0}{h}$.

Dem:

Completamente análoga a la anterior.

▼

7.5. Interpolación osculatoria. Polinomio de Hermite

En las anteriores secciones se ha abordado la resolución del problema de la reconstrucción de una función $f(x)$ a partir de una tabla de valores de la misma, $\{(x_i, f(x_i))\}_{i=1}^n$. En esta sección se aborda el estudio de como aproximar una función mediante interpolación polinómica cuando, además de la anterior tabla de valores, se conoce también el valor de la primera o sucesivas derivadas en los nodos.
Dentro de la interpolación osculatoria, uno de los casos más importantes es el siguiente: encontrar una función $f(x)$ de modo que ella y su derivada interpolen la tabla de valores $\{(x_i, y_i, y_i')\}_{i=0}^n$, esto es, $f(x_i) = y_i$, $f'(x_i) = y_i'$, $\forall i = 0, 1, \ldots, n$.
Para resolver este problema, se tiene en, primer lugar, que existe un único polinomio de grado menor o igual que $2n+1$ que interpola la tabla $\{(x_i, y_i, y_i')\}_{i=0}^n$; dicho polinomio se conoce como polinomio interpolante de Hermite $H_{2n+1}(x)$ para la tabla $\{(x_i, y_i, y_i')\}_{i=0}^n$.
La construcción del polinomio de Hermite puede abordarse por varios procedimientos, siendo los principales el de Lagrange y el de Newton, de los cuales únicamente analizaremos el enfoque lagrangiano, el cual se puede abordar del siguiente modo:

Teorema 7.5.1
Sean los nodos $x_0 < x_1 < x_2 < \cdots . < x_n$. Entonces existen unos únicos conjuntos de funciones $\{\Phi_i(x)\}_{i=0}^n$ y $\{\Psi_i(x)\}_{i=0}^n$, tales que:

1. $\Phi_i(x)$, $\Psi_i(x)$ son polinomios de grado $2n+1$, $\forall i = 0, 1, \ldots, n$.
2. $\Phi_i(x_i) = 1$, $\Phi_i'(x_i) = 0$, $\Psi_i(x_i) = 0$, $\Psi_i'(x_i) = 1$, $\forall i = 0, 1, \ldots, n$.
3. $\Phi_i(x_j) = \Phi_i'(x_j) = 0$, $\Psi_i(x_j) = \Psi_i'(x_j) = 0$, $\forall i, j = 0, 1, \ldots, n$, $i \neq j$.

Dem:
La función $\Phi_i(x)$ y su derivada se anulan en los nodos $x_0, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$. Por tanto, dichos nodos son raíz doble de $\Phi_i(x)$ y, por tanto, $\Phi_i(x)$ es divisible por el polinomio

$$(x - x_0)^2(x - x_1)^2 \cdots (x - x_{i-1})^2(x - x_{i+1})^2 \cdots (x - x_n)^2$$

y por ello, también lo es por $\ell_i^2(x)$ siendo $\ell_i(x)$ la i-ésima función de base de la interpolación de Lagrange. Puesto que $\Phi_i(x)$ debe ser un polinomio de grado $2n+1$, debe de cumplirse que

$$\Phi_i(x) = (\alpha_i x + \beta_i)\ell_i^2(x).$$

Los coeficientes α_i, β_i se determinan a partir de las condiciones $\Phi_i(x_i) = 1$ y $\Phi_i'(x_i) = 0$; así, se tiene que:

$$\Phi_i(x_i) = (\alpha_i x_i + \beta_i)\ell_i^2(x_i), \text{ y, por tanto, } \alpha_i x_i + \beta_i = 1.$$

Por otra parte, se tiene:

$$\Phi_i'(x) = \alpha_i \ell_i^2(x) + 2(\alpha_i x + \beta_i)\ell_i(x)\ell_i'(x)$$

de donde

$$\Phi_i'(x_i) = \alpha_i \ell_i^2(x_i) + 2(\alpha_i x_i + \beta_i)\ell_i(x_i)\ell_i'(x_i), \quad \text{y, por tanto, } \alpha_i + 2\ell_i'(x_i) = 0,$$

de donde se obtiene:

$$\alpha_i = -2\ell_i'(x_i), \quad \beta_i = 1 + 2\ell_i'(x_i)x_i,$$

y por tanto:

$$\alpha_i x + \beta_i = -2\ell_i'(x_i)x + 1 + 2\ell_i'(x_i) = 1 - 2\ell_i'(x_i)(x - x_i),$$

Así, $\Phi_i(x) = [1 - 2\ell_i'(x_i)(x - x_i)]\,\ell_i^2(x)$.
Para determinar $\Psi_i(x)$ procederemos de un modo similar. Así, puesto que la función $\Phi_i(x)$ y su derivada se anulan en los nodos $x_0, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$, y dado que ésta debe de ser un polinomio de grado $2n+1$, $\Psi_i(x)$, debe tener la forma:

$$\Psi_i(x) = (\gamma_i x_i + \delta_i)\ell_i^2(x)$$

Puesto que $\Psi_i(x_i) = 0$ se tiene que $\Psi_i(x)$ debe contener como factor a $x - x_i$ (el cual no está contenido en $\ell_i(x)$) y, por tanto, $\gamma_i x_i + \delta_i = C_i(x - x_i)$; de este modo se satisface

$$\Psi_i'(x) = C_i \ell_i^2(x) + 2C_i(x - x_i)\ell_i(x)\ell_i'(x),$$

y dado que $\Psi_i'(x_i) = 1$, se tiene

$$1 = C_i \ell_i^2(x) + 2C_i(x - x_i)\ell_i(x_i)\ell_i'(x_i) = C_i \cdots 1 + 0 = c_i.$$

De este modo, resulta que $\Psi_i(x) = (x - x_i)\ell_i^2(x)$.

▼

Teorema 7.5.2
El polinomio $H_{2n+1}(x) = \sum_{i=0}^{n} [\Phi_i(x)y_i + \Psi_i(x)y_i']$ interpola la tabla de valores de la función $f(x)$ y su derivada dada por $\{(x_i, y_i, y_i')\}_{i=0}^n$.

Dem:
Sea $x_j \in \{x_0, x_1, \ldots, x_n\}$, entonces:

$$H_{2n+1}(x_j) = \sum_{i=0}^{n} [\Phi_i(x_j)y_i + \Psi_i(x_j)y_i'] = \sum_{i=0}^{n} [\delta_{i,j}y_i + 0y_i'] = \delta_{j,j}y_j = y_j,$$
$$H_{2n+1}'(x_j) = \sum_{i=0}^{n} [\Phi_i'(x_j)y_i + \Psi_i(x_j)'y_i'] = \sum_{i=0}^{n} [0y_i + \delta_{i,j}y_i'] = \delta_{j,j}y_j' = y_j',$$

donde $\delta_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$ representa la delta de Kronecher. Por tanto, se verifica el teorema.

▼

7.6. Interpolación segmentaria

En un principio, puede parecer que obtener más precisión en las técnicas de interpolación es cuestión de tomar un mayor número de puntos en el intervalo donde se requiere interpolar; sin embargo, esto no siempre es así, basta tratar de aproximar la función $f(x) = \frac{1}{1+25x^2}$ en el intervalo $[-1,1]$ a partir de su tabla de valores en los puntos $x_0 = -1$, $x_k = x_0 + k \cdots h$, $k = 0, 1, \ldots, n$ con $h = \frac{2}{n}$ para observar que cuando aumenta n, el error lejos de disminuir, crece desmesuradamente alrededor de los puntos $x = -1$, $x = 1$. Este problema puede ser resuelto si en lugar de realizar la interpolación con un único polinomio interpolador para todo el intervalo [-1,1], se subdivide éste en subintervalos y se utiliza en cada uno de ellos un polinomio interpolante distinto. Para ello, definiremos en primer lugar el concepto de polinomio trazador (o splin) como:

Definición 7.6.1
Sea un intervalo $[a,b]$ en $\mathbb{R}$ y sea $a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$ una partición de $[a,b]$ en subintervalos. Diremos que $S(x)$ es un **polinomio trazador** o **splin** en $[a,b]$ asociado a la partición $a = x_0 < x_1 < \ldots . < x_{n-1} < x_n = b$ si $\exists S_0(x), S_1(x), \ldots ., S_{n-1}(x)$ polinomios de modo que

$$S(x) = \begin{cases} S_0(x) & x \in [x_0, x_1[\\ S_1(x) & x \in [x_1, x_2[\\ \ldots\ldots & \ldots\ldots\ldots \\ S_i(x) & x \in [x_i, x_{i+1}[\\ \ldots. & \ldots\ldots\ldots \\ S_{n-1}(x) & x \in [x_{n-1}, x_n] \end{cases},$$

Si además $\forall i = 0, 1, \ldots, n-1,\ S_i(x)$ es un polinomio de grado k y se cumplen las condiciones de continuidad

$$S_i(x_{i+1}) = S_{i+1}(x_{i+1}), \quad S_i'(x_{i+1}) = S_{i+1}'(x_{i+1}), \ldots ., S_i^{k-1)}(x_{i+1}) = S_{i+1}^{k-1)}(x_{i+1}),$$

se dice que $S(x)$ es un splin de orden k.

Definición 7.6.2
Diremos que un **splin** $S(x)$ **interpola la tabla de valores de una función** $\{(x_i, y_i)\}_{i=0}^n$ si $\forall i = 0, 1, \ldots, n$, se verifica que $S(x_i) = y_i$.

Nótese que los nodos de la tabla y de la partición en donde se define el splin no tienen necesariamente que coincidir.

Dentro de la interpolación por polinomios trazadores (splines) merecen especial consideración los llamados splines lineales y los splines cúbicos (de tercer orden). Respecto a los splines lineales se verifica:

Teorema 7.6.1
Dada la tabla de valores $\{(x_i, y_i)\}_{i=0}^n$ de una función $f(x)$, existe un único splin de grado uno, asociado a la partición $x_0 < x_1 < \cdots < x_n$ que interpola dicha función en dichos nodos.

Dem:

Sea $S(x)$ un splin de grado uno asociado a la partición

$$S(x) = \begin{cases} S_0(x) & x \in [x_0, x_1] \\ S_1(x) & x \in [x_1, x_2] \\ \ldots\ldots & \ldots\ldots\ldots \\ S_i(x) & x \in [x_i, x_{i+1}] \\ \ldots. & \ldots\ldots\ldots \\ S_{n-1}(x) & x \in [x_{n-1}, x_n] \end{cases}$$

donde por condición de continuidad se permite escribir los subintervalos como cerrados. Por interpolar la tabla $S_i(x)$ debe ser el único polinomio de grado uno que interpola $\{(x_i, y_i), (x_{i+1}, y_{i+1})\}$, y por tanto se verifica que

$$S_i(x) = \frac{x - x_{i+1}}{x_i - x_{i+1}} y_i + \frac{x - x_i}{x_{i+1} - x_i} y_{i+1}, \quad \forall i = 0, 1, \ldots, n-1,$$

con lo que se tiene el resultado.

▼

Respecto a la interpolación mediante splines cúbicos se verifica el siguiente resultado:

Teorema 7.6.2
Dada la tabla de valores $\{(x_i, y_i)\}_{i=0}^n$ de una función $f(x)$, existe un único splin de cúbico asociado a la partición $x_0 < x_1 < \cdots < x_n$ que interpola dicha función en dichos nodos y que satisface $S_0''(x_0) = 0$, y $S_{n-1}''(x_n) = 0$.

Dem:

Sea el splin $S(x)$ de grado tres, definido como

$$S(x) = \begin{cases} S_0(x) & x \in [x_0, x_1] \\ S_1(x) & x \in [x_1, x_2] \\ \ldots\ldots & \ldots\ldots\ldots \\ S_i(x) & x \in [x_i, x_{i+1}] \\ \ldots. & \ldots\ldots\ldots \\ S_{n-1}(x) & x \in [x_{n-1}, x_n] \end{cases}$$

que interpola la tabla $\{(x_i, y_i)\}_{i=0}^n$. Entonces $\forall i = 0, 1, \ldots n-1$, se tiene:

$$S_i(x_i) = y_i, \quad S_i(x_{i+1}) = y_{i+1},\ S_i'(x_{i+1}) = S_{i+1}'(x_{i+1}),\ S_i''(x_{i+1}) = S_{i+1}''(x_{i+1}),$$

expresiones en las que se recogen la condiciones de interpolación y continuidad. Por ser $S(x)$ un splin cúbico, las funciones $S_i(x)$ son polinomios de grado tres, y por tanto, sus primeras derivadas son polinomios de grado dos y sus segundas derivadas son polinomios de grado uno, por tanto, la segunda derivada $S_i''(x)$ puede determinarse conociendo su valor en los extremos.

Sea $S''_i(x_i) = z_i$, $S''_i(x_{i+1}) = z_{i+1}$, $\forall i = 1, 2, \ldots, n-1$. A partir de esto se puede obtener $S''_i(x)$ mediante interpolación lineal resultando:

$$S''_i(x) = \frac{x - x_{i+1}}{x_i - x_{i+1}} z_i + \frac{x - x_i}{x_{i+1} - x_i} z_{i+1}, \qquad \forall i = 0, 1, \ldots, n-1.$$

Definiendo $h_i = x_{i+1} - x_i$, $\forall i = 0, \ldots, n-1$ se obtiene

$$S''_i(x) = \frac{x_{i+1} - x}{h_i} z_i + \frac{x - x_i}{h_i} z_{i+1}, \qquad \forall i = 0, 1, \ldots, n-1,$$

a partir de lo cual obtenemos, integrando dos veces con respecto a la variable x:

$$S_i(x) = \frac{1}{6}\frac{(x_{i+1} - x)^3}{h_i} z_i + \frac{1}{6}\frac{(x - x_i)^3}{h_i} z_{i+1} + p_i(x), \quad \forall i = 0, 1, \ldots, n-1,$$

donde $p_i(x)$ es un polinomio de grado uno en x que expresaremos como $p_i(x) = \alpha_i(x_{i+1} - x) + \beta_i(x - x_i)$. Por tanto, $\forall i = 0, 1, \ldots, n-1$, se obtiene

$$S_i(x) = \frac{1}{6}\frac{(x_{i+1} - x)^3}{h_i} z_i + \frac{1}{6}\frac{(x - x_i)^3}{h_i} z_{i+1} + \alpha_i(x_{i+1} - x) + \beta_i(x - x_i).$$

Dicho polinomio debe interpolar la tabla $\{(x_i, y_i)\}_{i=0}^n$, por tanto, se verifica $S_i(x_i) = y_i$, $S_i(x_{i+1}) = y_{i+1}$ de donde se tiene inmediatamente

$$\frac{1}{6}h_i^2 z_i + \alpha_i h_i = y_i \quad \rightarrow \alpha_i = \frac{y_i}{h_i} - \frac{1}{6} z_i h_i$$
$$\frac{1}{6}h_i^2 z_{i+1} + \beta_i h_i = y_{i+1} \quad \rightarrow \beta_i = \frac{y_{i+1}}{h_i} - \frac{1}{6} z_{i+1} h_i.$$

Por tanto se tiene

$$S_i(x) = \frac{1}{6}\frac{(x_{i+1} - x)^3}{h_i} z_i + \frac{1}{6}\frac{(x - x_i)^3}{h_i} z_{i+1} + \left(\frac{y_i}{h_i} - \frac{1}{6} z_i h_i\right)(x_{i+1} - x) +$$
$$+ \left(\frac{y_{i+1}}{h_i} - \frac{1}{6} z_{i+1} h_i\right)(x - x_i) \qquad \forall i = 0, \ldots, n-1.$$

Por otra parte, se debe verificar la condición de continuidad para la primera derivada $S'(x)$; así, $S'_i(x_i + 1) = S'_{i+1}(x_i + 1)$, $i = 0, \ldots, n-2$. Derivando $S_i(x)$ y $S_{i+1}(x)$ se obtiene:

$$S'_i(x) = -\frac{1}{2}\frac{(x_{i+1} - x)^2}{h_i} z_i + \frac{1}{2}\frac{(x - x_i)^2}{h_i} z_{i+1} + \frac{1}{6} z_i h_i - \frac{y_i}{h_i} + \frac{y_{i+1}}{h_i} - \frac{1}{6} z_{i+1} h_i$$

$$S'_{i+1}(x) = -\frac{1}{2}\frac{(x_{i+2} - x)^2}{h_{i+1}} z_{i+1} + \frac{1}{2}\frac{(x - x_{i+1})^2}{h_{i+1}} z_{i+2} + \frac{1}{6} z_{i+1} h_{i+1} - \frac{y_{i+1}}{h_{i+1}} + \frac{y_{i+2}}{h_{i+1}} - \frac{1}{6} z_{i+2} h_{i+1},$$

y por tanto,

$$S'_i(x_{i+1}) = \frac{1}{2} h_i z_{i+1} + \frac{1}{6} z_i h_i - \frac{y_i}{h_i} + \frac{y_{i+1}}{h_i} - \frac{1}{6} z_{i+1} h_i$$

$$S'_{i+1}(x_{i+1}) = -\frac{1}{2}h_{i+1}z_{i+1} + \frac{1}{6}z_{i+1}h_{i+1} - \frac{y_{i+1}}{h_{i+1}} + \frac{y_{i+2}}{h_{i+1}} - \frac{1}{6}z_{i+2}h_{i+1},$$

igualando y operando se obtiene

$$\frac{1}{6}z_i h_i + \frac{1}{3}\left(h_i + h_{i+1}\right) z_{i+1} + \frac{1}{6}h_{i+1}z_{i+2} = \frac{y_i}{h_i} - \left[\frac{1}{h_i} + \frac{1}{h_{i+1}}\right] y_{i+1} + \frac{y_{i+2}}{h_{i+1}},$$

y finalmente

$$z_i h_{i+2}\left(h_i + h_{i+1}\right) z_{i+1} + h_{i+1}z_{i+2} = \frac{6}{h_{i+1}}\left(y_{i+2} - y_{i+1}\right) - \frac{6}{h_i}\left(y_{i+1} - y_i\right)$$

o lo que es equivalente:

$$z_{i-1}h_{i+1}\left(h_{i-1} + h_i\right) z_i + h_i z_{i+1} = \frac{6}{h_i}\left(y_{i+1} - y_i\right) - \frac{6}{h_{i-1}}\left(y_i - y_{i-1}\right).$$

Por tanto, y puesto que $z_0 = z_n = 0$, los valores $z_1, \ldots, z_{n-1}$ satisfacen el sistema de ecuaciones lineales

$$\begin{bmatrix} a_1 & h_1 & 0 & 0 & 0 & \ldots. & 0 & 0 & 0 \\ h_1 & a_2 & h_2 & 0 & 0 & \ldots. & 0 & 0 & 0 \\ 0 & h_2 & a_3 & h_3 & 0 & \ldots. & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \\ 0 & 0 & 0 & 0 & 0 & \ldots. & h_{n-2} & a_{n-2} & h_{n-1} \\ 0 & 0 & 0 & 0 & 0 & \ldots. & 0 & h_{n-1} & a_{n-1} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \ldots \\ z_{n-2} \\ z_{n-1} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \ldots\ldots \\ b_{n-2} \\ b_{n-1} \end{bmatrix}$$

donde:

$$a_i = 2\left(h_i + h_{i-1}\right), \quad b_i = v_i - v_{i-1}, \text{ con } v_i = \frac{6}{h_i}\left(y_{i+1} - y_i\right).$$

▼

Ejemplo 7.6.1
Construir el splin cúbico natural correspondiente a la tabla

n	0	1	2	3
x_n	1	2	0	3
y_n	3	2	-4	5

Para construir el splin cúbico natural correspondiente a la tabla procedemos, en primer lugar, a reordenar la misma

n	0	1	2	3
x_n	0	1	2	3
y_n	-4	3	2	5

El paso $h_i = x_{i+1} - x_i = 1$ es constante e igual a uno. El problema se reduce a encontrar los valores z_1, z_2 correspondientes a $S''(x_i)$, lo cual se consigue a partir de las relaciones

$$h_i = x_{i+1} - x_i, \quad u_i = 2(h_{i+1} + h_i)$$
$$b_i = 6/h_i(y_{i+1} - y_i), \quad v_i = b_1 - b_{i-1}.$$

Finalmente, el problema se resuelve a partir del sistema de ecuaciones

$$\begin{bmatrix} u_1 & h_1 \\ h_1 & u_2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}.$$

Así se tiene $h_1 = h_2 = 1$, $u_1 = u_2 = 4$, $b_0 = 42$, $b_1 = -6$, $b_2 = 18$,$v_1 = -48$,$v_2 = 24$, por tanto, se debe resolver el sistema:

$$\begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} -48 \\ 24 \end{bmatrix}$$

cuya solución viene dada por $z_1 = -\frac{72}{5}$, $z_2 = \frac{48}{5}$; por tanto el splin cúbico que interpola la tabla de datos viene dado por:

$$S(x) = \begin{cases} S_0(x) = -\frac{12}{5}x^3 + \frac{47}{5}x - 4 & \text{si } x \in [0,1] \\ S_1(x) = 4x^3 - \frac{96}{5}x^2 + \frac{143}{5}x - \frac{25}{5} & \text{si } x \in [1,2] \\ S_2(x) = -\frac{8}{5}x^3 + \frac{72}{5}x^2 - \frac{193}{5}x + \frac{172}{5} & \text{si } x \in [2,3]. \end{cases}$$

♦

Tema 8

Diferenciación e integración numérica

8.1. Introducción

En el presente capítulo se aborda el estudio de un conjunto de técnicas numéricas cuyo objeto es, por una parte, conseguir aproximaciones numéricas al valor de las derivadas de una determinada función; y por otra, determinar el valor de una integral definida con la mayor precisión posible.

El primer problema es, en general, de gran dificultad, pues como se muestra en el ejemplo, de la proximidad de dos funciones no se sigue la proximidad de sus derivadas. Sean las funciones $f_1(x) = x^2 + x$, y $f_2(x) = x^2 + x + \epsilon \text{sen} \frac{x}{\epsilon^2}$ donde ϵ es una pequeña cantidad positiva; estas funciones satisfacen, por una parte:

$$|f_1(x) - f_2(x)| = \left|\epsilon \text{sen} \frac{x}{\epsilon^2}\right| \leq \epsilon \left|\text{sen} \frac{x}{\epsilon^2}\right| \leq \epsilon.$$

Por otra parte, se verifica que $f_1'(x) = 2x + 1$, y $f_2'(x) = 2x + 1 + \frac{1}{\epsilon} \cos \frac{x}{\epsilon^2}$; así, en $x = 0$ se tiene que

$$|f_1(0) - f_2(0)| \leq \epsilon, \quad |f_1'(0) - f_2'(0)| = \frac{1}{\epsilon}.$$

La primera cantidad es pequeña cuando ϵ es pequeño, afirmación que no es posible hacer respecto a las derivadas pues cuando ϵ es pequeño $\frac{1}{\epsilon}$ es grande.

El problema de la integración numérica es, en principio, mucho más sencillo, pues si en un intervalo $[a, b]$ se verifica $|f_1(x) - f_2(x)| \leq \epsilon$, entonces se verifica

$$\left|\int_a^b f_1(x)dx - \int_a^b f_2(x)dx\right| \leq \int_a^b |f_1(x) - f_2(x)|\, dx \leq \int_a^f \epsilon dx = \epsilon(b - a).$$

El método que se seguirá para obtener aproximaciones a las derivadas e integrales de funciones estará basado en encontrar una aproximación a la función mediante interpolación polinomial, para a continuación efectuar la operación requerida sobre el polinomio interpolante.

En este capítulo, el tema de la derivación numérica se abordará someramente; mientras que dada su importancia, la integración numérica se tratará con mayor detalle.

8.2. Derivación numérica

Sea $f(x)$ una función continua y suficientemente derivable en un intervalo $[a, b]$ de la que conocemos sus valores en los nodos $a \leq x_0 < x_1 < \cdots < x_{n-1} < x_n \leq b$. A partir de la tabla $\{(x_i, f(x_i))\}_{i=0}^n$ se construye el polinomio interpolante para la misma, que en su forma de Lagrange resulta $L_n(x) = \sum_{i=0}^{n} \ell_i(x) f(x_i)$. A partir de esta expresión, resulta:

$$f(x) = L_n(x) + \frac{1}{(n+1)!} f^{(n+1)}(\xi_x) w(x), \quad \text{donde } w(x) = \prod_{i=0}^{n} (x - x_i).$$

Derivando se obtiene

$$f'(x) = L'_n(x) + \frac{1}{(n+1)!} f^{(n+1)}(\xi_x) w'(x) + \frac{1}{(n+1)!} w(x) \frac{d}{dx} f^{(n+1)}(\xi_x)$$

donde $w'(x) = \sum_{i=0}^{n} \prod_{\substack{j=0 \\ j \neq k}}^{n} (x - x_j)$. Esta expresión se simplifica en el caso de ser evaluada en los nodos, pues $w'(x_k) = \prod_{\substack{j=0 \\ j \neq k}}^{n} (x - x_j)$.

El error que se comete al tomar la aproximación $f'(x) \approx L'_n(x)$ viene dado por $\frac{1}{(n+1)!} f^{(n+1)}(\xi_x) w'(x) + \frac{1}{(n+1)!} w(x) \frac{d}{dx} f^{(n+1)}(\xi_x)$, expresión que se simplifica en el caso de evaluar la derivada sobre uno de los nodos, pues en ellos $w(x_i) = 0$ y, por tanto, en este caso el error viene dado por $\frac{1}{(n+1)!} f^{(n+1)}(\xi_{x_k}) w'(x_k)$. A partir de esta última expresión, en el caso de un mallado regular de nodos $x_0, x_1, \ldots, x_n$ donde $x_i = x_0 + ih$, el error cometido al interpolar en el nodo x_k viene dado por $\frac{1}{(n+1)!} f^{(n+1)}(\xi_{x_k}) k(k-1) \cdots 1(-1)(-2) \cdots (k-n) h^n$.

Por tanto, el error es de orden n en h.

8.2.1. Fórmulas mas usuales de derivación numérica

A continuación exponemos un conjunto de fórmulas de uso muy común en derivación numérica. En este apartado consideraremos que los nodos $x_0, x_1, \ldots, x_n$ están igualmente espaciados de modo que $x_i = x_0 + ih$. En primer lugar, se tiene

$$f'(x_i) \approx \frac{y_{i+1} - y_i}{h},$$

la cual procede de derivar el polinomio interpolante en los nodos x_i, x_{i+1}.

Para determinar el orden del error mediante un método alternativo desarrollamos y_{i+1} en serie de Taylor alrededor del punto x_i, así:

$$y_{i+1} = y_i + y'_i h + \frac{1}{2} f"(\xi) h^2$$

de donde

$$f'(x_i) = \frac{y_{i+1} - y_i}{h} - \frac{1}{2} f''(\xi) h,$$

por tanto, la aproximación es de orden uno en h.
Puede comprobarse que la aproximación dada por $f'(x_i) \approx \frac{y_i - y_{i-1}}{h}$, es también de primer orden en h.
Otra aproximación de gran utilidad es la siguiente:

$$f'(x_i) \approx \frac{y_{i+1} - y_{i-1}}{2n}.$$

Para evaluar su precisión consideramos los desarrollos

$$y_{i+1} = y_i + y_i' h + \frac{1}{2} y_i^{"} h^2 + \frac{1}{6} f^{3)}(\xi_1) h^3, \quad \xi_1 \in]x_i, x_{i+1}[,$$

$$y_{i-1} = y_i - y_i' h + \frac{1}{2} y_i^{"} h^2 - \frac{1}{6} f^{3)}(\xi_2) h^3, \quad \xi_1 \in]x_{i-1}, x_i[,$$

restando ambas expresiones y dividiendo por $2h$ se obtiene

$$\frac{y_{i+1} - y_{i-1}}{2h} = y_i' + \frac{1}{6} \left[f^{3)}(\xi_1) + f^{3)}(\xi_2) \right] h^2.$$

Si f es una función al menos C^3 se tiene que $\exists \xi \in [\text{mín}\{\xi_1, \xi_2\}, \text{máx}\{\xi_1, \xi_2\}]$ de modo que $f^{3)}(\xi) = \frac{1}{2} \left[f^{3)}(\xi_1) + f^{3)}(\xi_2) \right]$; así

$$y_i' = \frac{y_{i+1} - y_{i-1}}{2h} - \frac{1}{3} f^{3)}(\xi) h^2.$$

Esta fórmula es más precisa que la anterior pues es de segundo orden en h.
Otra fórmula de gran interés es la aproximación a la segunda derivada dada por

$$y_i^{''} \approx \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}.$$

Para determinar su error consideremos los desarrollos

$$y_{i+1} = y_i + y_i' h + \frac{1}{2} y_i^{''} h^2 + \frac{1}{6} y_i^{3)} h^3 + \frac{1}{24} f^{4)}(\xi_1) h^4, \quad \xi_1 \in]x_i, x_{i+1}[$$

$$y_{i-1} = y_i - y_i' h + \frac{1}{2} y_i^{''} h^2 - \frac{1}{6} y_i^{3)} h^3 + \frac{1}{24} f^{4)}(\xi_2) h^4, \quad \xi_1 \in]x_{i-1}, x_i[.$$

A partir de estas expresiones obtenemos

$$y_{i+1} - 2y_i + y_{i-1} = y_i^{''} h^2 + \frac{1}{24} \left[f^{4)}(\xi_1) + f^{4)}(\xi_2) \right] h^4,$$

de donde dividiendo por h^2 se obtiene

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = y_i^{''} + \frac{1}{24} \left[f^{4)}(\xi_1) + f^{4)}(\xi_2) \right] h^2,$$

y por tanto,

$$y_i^{''} = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - \frac{1}{24} \left[f^{4)}(\xi_1) + f^{4)}(\xi_2) \right] h^2.$$

Finalmente, procediendo como antes se obtiene

$$y_i^{''} = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - \frac{1}{12} f^{4)}(\xi) h^2.$$

Así, la aproximación resultante es de segundo orden en h.

Fórmulas de alto orden en los extremos

A partir de la interpolación de Newton con puntos igualmente espaciados, se obtiene el conjunto de fórmulas de orden n dado por

$$y'_i = \frac{1}{h}\left[\triangle - \frac{1}{2}\triangle^2 + \frac{1}{3}\triangle^3 + \cdots + \frac{(-1)^{n-1}}{n}\triangle^n\right] y_i$$

$$y'_i = \frac{1}{h}\left[\nabla + \frac{1}{2}\nabla^2 + \frac{1}{3}\nabla^3 + \cdots + \frac{1}{n}\nabla^n\right] y_i$$

$$y''_i = \frac{1}{h^2}\left[\triangle - \frac{1}{2}\triangle^2 + \frac{1}{3}\triangle^3 + \cdots + \frac{(-1)^{n-1}}{n}\triangle^n\right]^2 y_i$$

$$y''_i = \frac{1}{h^2}\left[\nabla + \frac{1}{2}\nabla^2 + \frac{1}{3}\nabla^3 + \cdots + \frac{1}{n}\nabla^n + \dots\right]^2 y_i,$$

y, en general, para orden m se verifica

$$y_i^{m)} = \frac{1}{h^n}\left[\triangle - \frac{1}{2}\triangle^2 + \frac{1}{3}\triangle^3 + \cdots + \frac{(-1)^{n-1}}{n}\triangle^n\right]^m y_i$$

$$y_i^{m)} = \frac{1}{h^n}\left[\nabla + \frac{1}{2}\nabla^2 + \frac{1}{3}\nabla^3 + \cdots + \frac{1}{n}\nabla^n\right]^m y_i.$$

Derivación mediante extrapolación

Un método particularmente útil para aumentar la precisión de las fórmulas de derivación numérica es el de extrapolación, el cual consiste en obtener aproximaciones a la derivada con dos pasos distintos, h_1 y h_2, a partir de lo cual se obtienen sendas aproximaciones $\hat{y}'(h_1)$ y $\hat{y}'(h_2)$ a la derivada $y'(x_i)$. Si la función $y(x)$ es suficientemente derivable, si además se cumple que $O(h_1) = O(h_2) = O(h_1 - h_2)$, y el método de derivación es de orden $O^k(h)$, se verificará que

$$y'(x_i) = \hat{y}'(h_1) + a_1 h_1^k + O^{k+1}(h_1)$$
$$y'(x_i) = \hat{y}'(h_2) + a_1 h_2^k + O^{k+1}(h_2)$$

$$y'(x_i)h_2^k = \hat{y}'(h_1)h_2^k + a_1 h_1^k h_2^k + O^{k+1}(h_1)h_2^k$$
$$y'(x_i)h_1^k = \hat{y}'(h_2)h_1^k + a_1 h_2^k h_1^k + O^{k+1}(h_2)h_1^k$$

de donde se tiene,

$$y'(x_i)\left(h_2^k - h_1^k\right) = \hat{y}'(h_1)h_2^k - \hat{y}'(h_1)h_2^k + O^{2k+1}(h_1),$$

y por tanto,

$$y'(x_i) = \frac{\hat{y}'(h_1)h_2^k - \hat{y}'(h_1)h_2^k}{h_2^k - h_1^k} + O^{k+1}(h_1),$$

fórmula que resulta de un orden mayor que la inicial.

El método de extrapolación es particularmente interesante cuando la fórmula de derivación inicial es de tipo central, pues en ese caso el desarrollo del error cometido al aproximar la derivada mediante la fórmula numérica sólo contiene potencias pares, por lo que cada vez que se aplica el método, el orden de la fórmula resultante aumenta en dos unidades.

Ejemplo 8.2.1
Obtener mediante extrapolación una fórmula de derivación cuyo error sea de cuarto orden a partir de $\hat{y}'_i(h) = \frac{y_{i+1}-y_{i-1}}{2h}$.

Para un paso h, se tiene

$$y'_i = \hat{y}'_i(h) - \frac{1}{3}f^{3)}(\xi)h^2 + O^4(h).$$

Para un paso $2h$:

$$y'_i = \hat{y}'_i(2h) - \frac{4}{3}f^{3)}(\xi)h^2 + O^4(h).$$

Multiplicando la primera expresión por cuatro y restando se tiene

$$3y'_i = 4\hat{y}'_i(h) - \hat{y}'_i(2h) + O^4(h),$$

y por tanto:

$$y'_i = \frac{4\hat{y}'_i(h) - \hat{y}'_i(2h)}{3} + O^4(h),$$

lo que conduce a la fórmula

$$y'_i \approx \frac{-y_{i+2} + 8y_{i+1} - 8y_{i-1} + y_{i-2}}{12h}.$$

♦

8.3. Integración numérica

El problema de la integración (o cuadratura) numérica trata de determinar el valor de una integral $\int_a^b f(x)dx$ con una precisión definida de antemano. El método que seguiremos para obtener dicho valor será aproximar la función $f(x)$ por un polinomio, para a continuación integrar dicho polinomio.
Sean los puntos $a \leq x_0 < x_1 < x_2 < \cdots . < x_{n-1} < x_n \leq b$ y sea la aproximación $L_n(x) = \sum_{i=0}^{n} \ell_i(x)f(x_i)$ a $f(x)$ mediante el polinomio de interpolación de Lagrange en los nodos $x_0, \ldots, x_n$. A continuación se proceder a efectuar la aproximación

$$\int_a^b f(x)dx \approx \int_a^b \sum_{i=0}^{n} \ell_i(x)f(x_i)dx = \sum_{i=0}^{n} \left\{ \int_a^b \ell_i(x)dx \right\} f(x_i) = \sum_{i=0}^{n} A_i f(x_i),$$

donde $A_i = \int_a^b \ell_i(x)dx$.

Cuando $x_0 = a$ y $x_n = b$ se dice que la fórmula de cuadratura es cerrada. En caso de ser $a < x_0$ y $x_n < b$ se dice que la fórmula es abierta. En general, son más difíciles de obtener las fórmulas abiertas que las cerradas, si bien las primeras pueden ser construidas de modo que resulten de mayor precisión para la misma cantidad de nodos.

Diremos que una fórmula de cuadratura numérica es de orden n si es exacta para polinomios de grado menor o igual a n. La elección del método de interpolación nos asegura que tomando $n+1$ nodos $x_0, x_1, \ldots, x_n$ en un intervalo $[a, b]$, la cuadratura numérica resultante es al menos de orden n, pues si $f(x)$ es un polinomio de grado menor o igual que n se verifica que $L_n(x) = f(x)$, de donde se tiene la igualdad de resultados.

8.3.1. Fórmulas de Newton-Cotes

Las fórmulas de Newton-Cotes son fórmulas cerradas construidas del modo siguiente.
Sea $n > 0$, y sea $x_0 = a$, $h = \frac{b-a}{n}$, $x_k = x_0 + kh, \quad k = 0, 1, \ldots, n$. Sea $L_n(x)$ el polinomio de interpolación de Lagrange sobre los nodos $x_0, x_1, \ldots, x_n$. Para obtener la correspondiente fórmula de cuadratura bastará obtener los valores A_i, para lo cual efectuaremos el cambio de variable $x = x_0 + th$, así $dx = hdt$ y $x_i = x(i),\ i = 0, 1, \ldots, n$. De este modo

$$A_i = \int_a^b \ell_i(x)dx = \int_0^n \ell_i(x(t))hdt = h\int_0^n \ell_i(x(t))dt = \frac{b-a}{n}\int_0^n \ell_i(x(t))dt.$$

Por otra parte,

$$\ell_i(x) = \frac{(x-x_0)\ldots.(x-x_{i-1})(x-x_{i+1})\ldots.(x-x_n)}{(x_i-x_0)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)},$$

y dado que

$$x - x_k = x_0 + th - (x_0 + kh) = (t-k)h$$
$$x_i - x_k = x_0 + ih - (x_0 + kh) = (i-k)h,$$

se tiene

$$\ell_i(x) = \frac{th\ldots.(t-i+1)h\cdot(t-i-1)h\ldots.(t-n)h}{ih\ldots.h(-h)\ldots.(i-n)h} =$$
$$= (-1)^{n-i}\frac{t(t-1)\ldots(t-i+1)(t-i-1)\ldots.(t-n)}{i!(n-i)!}.$$

De este modo:

$$A_i = \frac{b-a}{n}\alpha_i, \quad \alpha_i = (-1)^{n-i}\int_0^n \frac{t(t-1)\cdots(t-i+1)(t-i-1)\cdots(t-n)}{i!(n-i)!}dt,$$

donde los coeficientes α_i son independientes del intervalo $[a, b]$. De este modo, es posible rescribir la fórmula de cuadratura numérica como

$$\int_a^b f(x)dx = \frac{b-a}{n}\sum_{i=0}^{n}\alpha_i f(x_i)$$

obteniéndose una fórmula que depende del intervalo únicamente a través del factor $b-a$ y de la posición de los nodos, pero no de los factores α_i.
A continuación obtendremos de modo explícito las cuadraturas numéricas para el caso particular $n = 1$ y $n = 2$.

Fórmula del trapecio

La cuadratura de Newton-Cotes con $n = 1$ recibe el nombre de fórmula del trapecio. Para determinar la fórmula del trapecio bastará determinar sus coeficientes α_0, α_1. Así:

$$\alpha_0 = (-1)^{1-0}\int_0^1 \frac{t-1}{0!1!}dt = \int_0^1 (t-1)dt = -\left[\frac{t^2}{2} - t\right]_0^1 = -\left[\frac{1}{2} - 1\right] = \frac{1}{2}$$

$$\alpha_1 = (-1)^{1-1}\int_0^1 \frac{t}{1!0!}dt = \int_0^1 t dt = \left[\frac{t^2}{2}\right]_0^1 = \frac{1}{2},$$

De este modo se tiene

$$T(f(x), a, b) = (b-a)\left[\frac{1}{2}f(x_0) + \frac{1}{2}f(x_1)\right],$$

o lo que es equivalente

$$T(f(x), a, b) = \frac{b-a}{2}\left[f(a) + f(b)\right].$$

Fórmula de Simpson

A continuación se aborda el estudio de la fórmula de Newton-Cotes de orden 2 o fórmula de Simpson. Al igual que antes deberemos obtener los valores de α_0, α_1, y α_2; así:

$$\alpha_0 = (-1)^{2-0}\int_0^2 \frac{(t-1)(t-2)}{0!2!}dt = \frac{1}{2}\int_0^2 (t^2 - 3t + 2)dt = \frac{1}{2}\left[\frac{t^3}{3} - \frac{3}{2}t^2 + 2t\right]_0^2 =$$
$$= \frac{1}{2}\left(\frac{8}{3} - \frac{12}{3} + 2\right) = \frac{1}{3},$$

$$\alpha_1 = (-1)^{2-1}\int_0^2 \frac{t(t-2)}{1!1!}dt = -\int_0^2 (t^2 - 2t)dt = -\left[\frac{t^3}{3} - t^2\right]_0^2 = -\left(\frac{8}{3} - 4\right) = \frac{4}{3},$$

$$\alpha_2 = (-1)^{2-2}\int_0^2 \frac{t(t-1)}{2!0!}dt = \frac{1}{2}\int_0^2 (t^2 - t)dt = \frac{1}{2}\left[\frac{t^3}{3} - \frac{t^2}{2}\right]_0^2 = \frac{1}{2}\left(\frac{8}{3} - 2\right) = \frac{1}{3},$$

de donde resulta

$$S(f(x), a, b) = \frac{b-a}{2}\left[\frac{1}{3}f(x_0) + \frac{4}{3}f(x_1) + \frac{1}{3}f(x_2)\right],$$

o lo que es equivalente

$$S(f(x), a, b) = \frac{b-a}{6}\left[f(a) + 4f(\frac{a+b}{2}) + f(b)\right].$$

.

8.4. Fórmulas compuestas

Para aumentar la precisión de las fórmulas de cuadratura numérica de Newton-Cotes una solución consiste en dividir el intervalo $[a,b]$ en subintervalos mas pequeños, aplicar en cada uno de ellos una fórmula de cuadratura numérica y sumar los resultados; así se tienen las fórmulas compuestas del trapecio y de Simpson.

8.4.1. Fórmula del trapecio compuesta

Sea la integral $\int_a^b f(x)dx$, y sea $n > 1$. Sea $h = \frac{b-a}{n}$ y sea $x_k = x_0 + k\,h$ donde $x_0 = a$, $k = 0, 1, \ldots, n$. Entonces, se tiene que

$$\int_a^b f(x)dx = \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} f(x)dx \approx \sum_{k=0}^{n-1} \frac{x_{k+1} - x_k}{2} \left[f(x_k) + f(x_{k+1})\right],$$

y por tanto:

$$\int_a^b f(x)dx \approx \frac{h}{2}\left[f(x_0) + 2\sum_{k=1}^{n-1} f(x_k) + f_(x_n)\right],$$

lo que es equivalente a

$$\int_a^b f(x)dx \approx \frac{b-a}{2n}\left[f(a) + 2\sum_{k=1}^{n-1} f\left(a + k\frac{b-a}{n}\right) + f(b)\right].$$

lo que constituye la fórmula del trapecio compuesta.

Ejemplo 8.4.1
Aproximar mediante la regla del trapecio compuesto con $n = 2$ y $n = 4$ el valor de la integral $\int_0^1 \operatorname{sen}(x^2)dx$. Utilizar cuatro decimales en los cálculos.

1. Para $n = 2$ se tiene: $h = 0.5$, $x_0 = 0$, $x_1 = 0.5$, $x_2 = 1$, por tanto:

$$\int_0^1 \operatorname{sen}(x^2)dx \approx \frac{1-0}{4}\left[\operatorname{sen} 0^2 + 2\operatorname{sen} 0.5^2 + \operatorname{sen} 1^2\right]$$

esto es,

$$\int_0^1 \operatorname{sen}(x^2)dx \approx 0.25(0+2\operatorname{sen} 0.25+\operatorname{sen} 1) = 0.25(0+2\cdot 0,2474+0.8415)$$

$$\int_0^1 \operatorname{sen}(x^2)dx \approx 0.3341\,.$$

2. Para $n = 4$ se tiene: $h = 0.25$, $x_0 = 0$, $x_1 = 0.25$, $x_2 = 0.5$, $x_3 = 0.75$, y $x_4 = 1$, por tanto:

$$\int_0^1 \operatorname{sen}(x^2)dx \approx \frac{1-0}{8}\left[\operatorname{sen} 0^2 + 2\operatorname{sen} 0,25^2 + 2\operatorname{sen} 0.5^2 + 2\operatorname{sen} 0.75^2 + \operatorname{sen} 1^2\right].$$

esto es

$$\int_0^1 \text{sen}\,(x^2)dx \approx 0.25(0{+}2\,\text{sen}\,0.0625{+}2\text{sen}\,0.25{+}2\,\text{sen}\,0.5625{+}\text{sen}\,1) = 0.3160\,.$$

El valor verdadero de la integral con cuatro cifras es

$$\int_0^1 \text{sen}\,(x^2)dx = 0.3103\,.$$

♦

8.4.2. Fórmula del Simpson compuesta

Sea la integral $\int_a^b f(x)dx$, y sea $n > 1$, y sea $h = \frac{b-a}{2n}$ sea $x_k = x_0 + k\,h$ donde $x_0 = a$, $k = 0, 1, \ldots, 2n$. Entonces se tiene

$$\int_a^b f(x)dx = \sum_{k=0}^{n-1} \int_{x_{2k}}^{x_{2k+2}} f(x)dx \approx \sum_{k=0}^{n-1} \frac{x_{2k+2} - x_{2k}}{6} \left[f(x_{2k}) + 4f(x_{2k+1}) + f(x_{2k+2})\right],$$

y por tanto:

$$\int_a^b f(x)dx = \frac{b-a}{6n} \left[f(x_0) + 4\sum_{k=0}^{n-1} f(x_{2k+1}) + 2\sum_{k=1}^{n-1} f(x_{2k}) + f(x_{2n})\right],$$

lo que constituye la fórmula de Simpson compuesta.

Ejemplo 8.4.2
Evaluar mediante la fórmula de Simpson compuesta con $n = 2$ la integral $\int_0^1 \text{sen}\,(x^2)dx$. Utilizar cuatro decimales en los cálculos.

Para $n = 2$, se tiene: $h = 0.25$, $x_0 = 0$, $x_1 = 0.25$, $x_2 = .5$, $x_3 = .75$, $x_4 = 1.00$, por tanto:

$$\int_0^1 \text{sen}\,(x^2)dx \approx \frac{1-0}{12} \left[\text{sen}\,0^2 + 4\,\text{sen}\,0,25^2 + 4\,\text{sen}\,0.75^2 + 2\,\text{sen}\,0.5^2 + \text{sen}\,1^2\right],$$

esto es

$$\int_0^1 \text{sen}\,(x^2)dx \approx 0,0833(0{+}4\,\text{sen}\,0.0625{+}4\,\text{sen}\,0.5625{+}2\,\text{sen}\,0.25{+}\text{sen}\,1) = 0.3098\,.$$

Nótese que este resultado ha requerido la evaluación de la función $f(x) = \text{sen}\,(x^2)$ en cinco puntos, las mismas que en el caso del trapecio compuesto con $n = 4$, siendo en este caso, mayor la precisión alcanzada. ♦

8.5. Cuadraturas gaussianas

Las fórmulas de Newton-Cotes forman parte de las denominadas fórmulas cerradas; estas fórmulas evalúan la función sobre un conjunto de nodos igualmente espaciados.

Si se eliminan las restricciones $x_0 = a$, $x_n = b$, y $x_k = x_0 + k\frac{b-a}{n}$, es posible encontrar fórmulas de cuadratura numérica de la forma

$$\int_a^b f(x)dx = \sum_{k=0}^{n} \gamma_k f(x_k)$$

de modo que sean exactas para polinomios de hasta grado $2n-1$. Para lograr tal fin se requiere que se satisfagan las igualdades

$$\int_a^b n^k dx = \sum_{k=0}^{n} \gamma_k x^k,\ k = 0, 1, \ldots, 2n-1,$$

lo que equivale a que se satisfagan las igualdades

$$\sum_{k=0}^{n} \gamma_k x^k = \frac{(b-a)^{k+1}}{k+1},\ k = 0, 1, \ldots, 2n-1.$$

Para lograr este propósito puede resolverse directamente el sistema anterior, obteniéndose de él los valores de γ_i y de x_i.

Ejemplo 8.5.1
Encontrar una fórmula de cuadratura numérica de la forma

$$\int_a^b f(x)dx = (b-a)\gamma f(x_0)$$

exacta para polinomios de grado uno.

La fórmula debe ser exacta para $f(f) = 1$ y para $f(x) = x$. Así, se tiene que

$$\gamma(b-a) = (b-a), \quad \gamma(b-a)x_0 = \frac{b^2-a^2}{2}.$$

De la primera de dichas ecuaciones se tiene que $\gamma = 1$, y de la segunda $(b-a)x_0 = \frac{b^2-a^2}{2} = \frac{(b-a)(b+a)}{2}$ de donde $x_0 = \frac{a+b}{2}$.

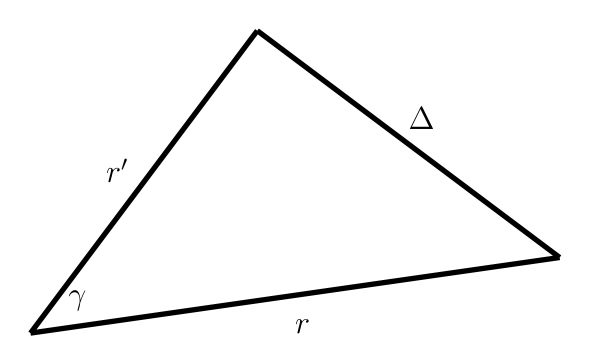
Así, la fórmula

$$\int_a^b f(x)dx \approx (b-a)f\left(\frac{a+b}{2}\right)$$

sólo requiere una evaluación de la función y es exacta para polinomios hasta grado uno. ♦

8.5.1. Polinomios de Legendre

El método seguido en el ejemplo anterior se complica extraordinariamente al crecer n. Por ello es preferible seguir una vía alternativa, la cual está basada en las propiedades de los llamados polinomios de Legendre. Los polinomios de Legendre aparecen al tratar de obtener el valor del inverso de un lado Δ de un triángulo en función de los otros dos, de longitudes r, r' y del ángulo γ comprendido entre ellos.



En el citado triángulo se satisface la relación

$$\Delta^2 = r^2 + r'^2 - 2rr' \cos\gamma,$$

y por tanto:

$$\frac{1}{\Delta} = \frac{1}{\sqrt{r^2 + r'^2 - 2rr' \cos\gamma}}.$$

Sea $r < r'$ y sea $t = \frac{r}{r'}$, $x = \cos\gamma$. Entonces se verifica que

$$\frac{1}{\Delta} = \frac{1}{r'} \frac{1}{\sqrt{1 + t^2 - 2tx}}$$

La función

$$G(t, x) = \frac{1}{\sqrt{1 + t^2 - 2tx}},$$

para un x fijo $-1 \leq x \leq 1$, puede desarrollarse en serie de potencias de t como

$$G(t, x) = \sum_{n=0}^{\infty} t^n P_n(x),$$

donde los $P_n(x)$ son los coeficientes de Taylor del desarrollo, los cuales vienen dados por

$$P_n(x) = \frac{1}{n!} \frac{d^n}{dt^n} \left[\frac{1}{\sqrt{1 + t^2 - 2tx}} \right] \Bigg|_{t=0}.$$

Definición 8.5.1
Llamamos **polinomios de Legendre** a los coeficientes del desarrollo en serie de Taylor respecto a la variable t alrededor de $t = 0$, denotándose por $P_n(x)$ al coeficiente de t^n en dicho desarrollo. La función $G(t, x)$ se denomina **función generatriz** de los polinomios de Legendre.

Teorema 8.5.1
Las funciones $P_n(x)$ son polinomios de grado n dados por $P_0(x) = 0$, $P_1(x) = x$, $(n+1)P_{n+1}(x) - (2n+1)xP_n(x) + nP_{n-1}(x) = 0$, $n \geq 1$.

Dem:
Para $n = 0$ se tiene que $P_0(x) = G(0, x) = 1$. Para $n = 1$ se verifica

$$P_1(x) = \frac{d}{dt}\left[\frac{1}{\sqrt{1+t^2-2tx}}\right]\Bigg|_{t=0} = -\frac{1}{2}\,\frac{2t-2x}{\left(\sqrt{1+t^2-2tx}\right)^3}\Bigg|_{t=0} = x.$$

Para obtener los valores de $P_n(x)$ con $n > 1$, se deriva con respecto a t la igualdad

$$\frac{1}{\sqrt{1+t^2-2tx}} = \sum_{n=0}^{\infty} t^n P_n(x)$$

obteniéndose

$$\frac{x-t}{(\sqrt{1+t^2-2tx})^3} = \sum_{n=1}^{\infty} nt^{n-1}P_n(t),$$

lo que es equivalente a

$$\frac{x-t}{1+t^2-2tx}G(t,x) = \sum_{n=1}^{\infty} nt^{n-1}P_n(t),$$

de donde

$$(x-t)\sum_{n=0}^{\infty} t^n P_n(x) = (1+t^2-2tx)\sum_{n=1}^{\infty} nt^{n-1}P_n(t),$$

$$\sum_{n=0}^{\infty} t^n xP_n(x) - \sum_{n=0}^{\infty} t^{n+1}P_n(x) = \sum_{n=1}^{\infty} nt^{n-1}P_n(t) + \sum_{n=1}^{\infty} nt^{n+1}P_n(t) - \sum_{n=1}^{\infty} 2nt^n xP_n(x),$$

lo que es equivalente a

$$\sum_{n=0}^{\infty} t^n xP_n(x) - \sum_{n=1}^{\infty} t^n P_{n-1}(x) =$$

$$= \sum_{n=0}^{\infty}(n+1)t^n P_{n+1}(t) + \sum_{n=2}^{\infty}(n-1)t^n P_{n-1}(t) - \sum_{n=1}^{\infty} 2nt^n xP_n(x).$$

Así, para $n = 0$ se verifica $P_1(x) = xP_0(x)$, para $n = 1$, $2P_2(x) = 3P_1(x) - P_0(x)$ y, para $n > 1$:

$$(n+1)P_{n+1} - (2n+1)xP_n(x) + nP_{n-1}(x) = 0, \; n > 1,$$

relación que también se verifica, tal como se ha visto anteriormente para $n = 1$.

A partir de estas relaciones se tiene de modo inmediato que $P_n(x)$ es un polinomio de grado n. ▼

Definición 8.5.2
Se dice que un conjunto de **funciones** $\{f_n(x)\}_{n\in I}$ son **ortogonales** en un intervalo $[a,b]$ con **respecto a una función de peso** $\rho(x)>0,\ x\in]a,b[$ si $\forall n,m\in I,\ n\neq m$, se verifica la igualdad $\int_a^b f_n(x)f_m(x)\rho(x)dx=0$.

Ejemplo 8.5.2
Los polinomios de Tchevichev $T_n(x)$ son ortogonales en $[-1,1]$ con respecto al peso $\rho=\frac{1}{\sqrt{1-x^2}}$, pues si $n\neq m$ se verifica que

$$\int_{-1}^{1} T_n(x)T_m(x)\frac{dx}{\sqrt{1-x^2}}=0,$$

donde consideraremos $n>m$.

En efecto, sea $x=\cos t$, entonces, se tiene que $t=\arccos x$ de donde $dt=\frac{-dx}{\sqrt{1-x^2}}$, y $t\in[0,\pi]$. Por otra parte:

$$T_n(x)=T_n(\cos t)=\cos(n\arccos(\cos t))=\cos nt,$$

así se tiene que

$$\int_{-1}^{1} T_n(x)T_m(x)\frac{dx}{\sqrt{1-x^2}}=\int_{1}^{-1} T_n(x)T_m(x)\frac{-dx}{\sqrt{1-x^2}}=\int_0^\pi \cos nt\cos mt dt=$$

$$=\int_0^\pi \frac{1}{2}[\cos(n+m)t+\cos(n-m)x]dx=$$

$$=\frac{1}{2(n+m)}\text{sen}\,(n+m)x\Big|_0^\pi+\frac{1}{2(n-m)}\text{sen}\,(n-m)x\Big|_0^\pi=0.$$

♦

Teorema 8.5.2
Los polinomios de Legendre son ortogonales en el intervalo $[-1,1]$ con respecto al peso $\rho(x)=1$.

Dem:

Considérese el producto de funciones generatrices

$$G(t,x)G(s,x)=\sum_{n=0}^{\infty}P_n(x)t^n\sum_{m=0}^{\infty}P_m(x)s^m.$$

Por tanto:

$$G(t,x)G(s,x)=\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}P_n(x)P_m(x)t^ns^m.$$

Integrando ambos miembros con respecto a x en $[-1,1]$ se tiene

$$\int_{-1}^{1}G(t,x)G(s,x)dx=\int_{-1}^{1}\left[\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}P_n(x)P_m(x)t^ns^m\right]dx,$$

de donde

$$\int_{-1}^{1} G(t,x)G(s,x)dx = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\left[\int_{-1}^{1}[P_n(x)P_m(x)dx]\right]t^n s^m.$$

La primitiva correspondiente a la integral del primer miembro viene dada por

$$\int_{-1}^{1} G(t,x)G(s,x)dx = -\frac{1}{\sqrt{t\,s}}\log[\sqrt{s(1+t^2-2t\,x)}+\sqrt{t(1+s^2-2s\,x)}],$$

y por tanto:

$$\int_{-1}^{1} G(t,x)G(s,x)dx = \frac{1}{\sqrt{t\,s}}\log\frac{(1+t)\sqrt{s}+(1+s)\sqrt{t}}{(1-t)\sqrt{s}+(1-s)\sqrt{t}},$$

de donde

$$\int_{-1}^{1} G(t,x)G(s,x)dx == \frac{1}{\sqrt{t\,s}}\log\frac{(\sqrt{t}+\sqrt{s})(1+\sqrt{t\,s}}{(\sqrt{t}+\sqrt{s})(1-\sqrt{t\,s}}.$$

Finalmente,

$$\int_{-1}^{1} G(t,x)G(s,x)dx = \frac{1}{\sqrt{t\,s}}\log\frac{1+\sqrt{t\,s}}{1-\sqrt{t\,s}}$$

Sea la función $f(z)=\frac{1}{z}\log\frac{1+z}{1-z}$. Si $|z|<1$ se tiene que

$$\frac{1}{z}\log\frac{1+z}{1-z} = 2\sum_{n=0}^{\infty}\frac{z^{2n}}{2n+1},$$

y por tanto:

$$\int_{-1}^{1} G(t,x)G(s,x)dx = 2\sum_{n=0}^{\infty}\frac{t^n s^n}{2n+1}.$$

El desarrollo asintótico de una función es único; por tanto, para $n \neq m$, debe verificarse

$$\int_{-1}^{1} P_n(x)P_m(x)dx = 0.$$

▼

Teorema 8.5.3
Sea $P_n(x)$ el polinomio de Legendre de grado n. Entonces $\int_{-1}^{1}[P_n(x)]^2\,dx = \frac{2}{2n+1}$.

Dem:

En el teorema anterior se tiene que los desarrollos asintóticos también deben coincidir cuando $n=m$; así, se tiene que $\int_{-1}^{1}[P_n(x)]^2\,dx = \frac{2}{2n+1}$. ▼

Teorema 8.5.4
Sea $P_n(x)$ el polinomio de Legendre de grado n y sea $Q(x)$ un polinomio de grado $m<n$. Entonces $\int_{-1}^{1} Q(x)P_n(x)=0$.

Dem:

Los polinomios de Legendre $\{P_0(x), P_1(x), \ldots, P_m(x)\}$ constituyen una base del espacio vectorial de los polinomios de grado menor o igual que m. Por tanto, existen escalares $\alpha_0, \ldots, \alpha_m$ de modo que $Q(x) = \sum_{i=0}^{m} \alpha_1 P_i(x)$. De este modo, se tiene que

$$\int_{-1}^{1} Q(x)P_n(x) = \int_{-1}^{1} \sum_{i=0}^{m} \alpha_i P_i(x)P_n(x)dx = \sum_{i=0}^{m} \alpha_i \int_{-1}^{1} P_i(x)P_n(n)dx = 0,$$

pues dado que $i < m$, por 8.5.1 se tiene que $\int_{-1}^{1} P_i(x)P_n(n)dx = 0$ de donde se sigue el resultado. ▼

Teorema 8.5.5
Sean los nodos $-1 \leq x_1 < x_2 < \cdots < x_n \leq b$ y sea $P_n(x)$ el polinomio de Legendre de grado n. Entonces, el conjunto

$$B = \{\ell_1(x), \ldots, \ell_n(x), P_n(x), xP_n(x), \ldots, x^{n-1}P_n(x)\},$$

donde $\{\ell_1(x), \ldots, \ell_n(x)\}$ son las funciones de base de la interpolación de Lagrange sobre los nodos $x_1, \ldots, x_n$ constituye una base del espacio vectorial de los polinomios de grado menor o igual que $2n - 1$.

Dem:
Sea $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ y sea $\beta_0, \ldots, \beta_{n-1} \in \mathbb{R}$ de modo que

$$\sum_{i=1}^{n} \alpha_i \ell_i(x) + \sum_{i=0}^{n-1} \beta_1 x^i P_n(x) = 0.$$

En la expresión anterior el único término de grado $2n-1$ procede de $\beta_{n-1}x^{n-1}P_n(x)$. Por tanto, $\beta_{n-1} = 0$. Supongamos que $\beta_{n-1} = \cdots = \beta_{n-k} = 0$ con $k < n$. Entonces,

$$\sum_{i=1}^{n} \alpha_i \ell_i(x) + \sum_{i=0}^{n-1} \beta_1 x^i P_n(x) = \sum_{i=1}^{n} \alpha_i \ell_i(x) + \sum_{i=0}^{n-k-1} \beta_1 x^i P_n(x) = 0.$$

El término de mayor grado $2n - k - 1$ únicamente puede proceder de $x^{n-k-1}P_n(x)$ por tanto, $\beta_{n-k-1} = 0$. De este modo se tiene que $\beta_i = 0, i = 0, \ldots, n-1$. Entonces, se tiene que $\sum_{i=1}^{n} \alpha_i \ell_i(x) = 0$, pero en este caso se tiene que $\alpha_i = 0$ pues el conjunto de funciones de base de la interpolación de Lagrange es linealmente independiente.

Finalmente, puesto que el cardinal del conjunto B es igual a la dimensión del espacio vectorial, la cual es igual a $2n$, se tiene que B es una base del espacio vectorial de los polinomios de grado menor o igual que $2n - 1$. ▼

8.5.2. Fórmulas de cuadratura de Gauss

A partir del teorema anterior se puede construir un conjunto de fórmulas de cuadratura numérica en el intervalo $[a, b]$, llamadas cuadraturas gaussianas, del siguiente modo:

Sea $P_n(x)$ el polinomio de Legendre de grado $n > 1$ y sean $-1 < x_1 < x_2 < \cdots < x_n < 1$ sus raíces. Sea $f : [-1, 1] \to \mathbb{R}$ una función y $L_n(x) = \sum_{n=1}^{\infty} \ell_i(x) f(x_i)$ su polinomio de interpolación construido sobre los nodos $x_1, x_2, \ldots, x_n$

$$G(f, n) = \int_{-1}^{1} L_n(x)dx = \sum_{i=1}^{n} \alpha_i f(x_i), \quad \alpha_i = \int_{-1}^{1} \ell_i(x)dx,$$

lo que constituye la fórmula de integración gaussiana de n nodos.

Teorema 8.5.6
La fórmula de integración gaussiana de n es exacta para polinomios hasta grado $2n - 1$.

Dem:

Sea $Q_k(x)$ un polinomio de grado k menor que $2n - 1$. Sean $x_1 < \cdots < x_n$ las raíces del polinomio de Legendre de grado n y sea $\{\ell_i(x)\}_{i=1}^{n}$ el conjunto de funciones de base de la interpolación de Lagrange sobre los nodos $x_1, \ldots, x_n$, y sea $P_n(x)$ el polinomio de Legendre de grado n.

Puesto que el conjunto $\{\ell_1(x), \ldots, \ell_n(x), P_n(x), xP_n(x), \ldots, x^{n-1}P_n(x)\}$ constituye una base del espacio vectorial de los polinomios de grado menor o igual que $2n - 1$, se tiene que existe un único conjunto de coeficientes $\gamma_1, \ldots, \gamma_{2n}$ de modo que

$$Q_k(x) = \sum_{i=1}^{n} \gamma_i \ell_i(x) Q_k(x_i) + \sum_{i=1}^{n} \gamma_{n+i} x^{i-1} P_n(x).$$

Integrando ambos miembros se tiene

$$\int_{-1}^{1} Q_k(x)dx = \int_{-1}^{1} \left[\sum_{i=1}^{n} \gamma_i \ell_i(x) Q_k(x_i) + \sum_{i=1}^{n} \gamma_{n+i} x^{i-1} P_n(x) \right] dx.$$

por otra parte se tiene que

$$Q_k(x_j) = \sum_{i=1}^{n} \gamma_i \ell_i(x_j) Q_k(x_i) + \sum_{i=1}^{n} \gamma_{n+i} x_j^{i-1} P_n(x_j) = \gamma_j,$$

pues $\ell_i(x_j) = \delta_{i,j}$ donde $\delta_{i,j}$ es la δ de Kronecher y $P_n(x_j) = 0$ pues los nodos son las raíces de $P_n(x)$. De este modo, se tiene que

$$\int_{-1}^{1} Q_k(x)dx = \sum_{i=1}^{n} \alpha_i Q_k(x_i) + \sum_{i=0}^{n} \int_{-1}^{1} x_{i-1} P_n(x)dx.$$

Por otra parte, se tiene por 8.5.1 que $\int_{-1}^{1} x_{i-1} P_n(x) dx = 0$, de donde

$$\int_{-1}^{1} Q_k(x)dx = \sum_{i=1}^{n} \alpha_i Q_k(x_i) = G(Q_k(x), n).$$

▼

Cuando la integración se realiza en un intervalo arbitrario $[a, b]$, la integración puede reducirse al intervalo $[-1, 1]$ mediante la transformación $x = \frac{a-b}{2}t + \frac{a+b}{2}$. De este modo se tiene que $dx = \frac{b-a}{2}dt$. Así

$$\int_{a}^{b} f(x)dx = \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{a-b}{2}t + \frac{a+b}{2}\right)dt,$$

por tanto, si $t_1, t_2, \ldots, t_n$ son las raíces del polinomio de Legendre de grado n, se tiene la fórmula de cuadratura gaussiana:

$$G(f, a, b, n) = \frac{b-a}{2}\sum_{i=1}^{n} \alpha_i f(\left(\frac{a-b}{2}t_i + \frac{a+b}{2}\right), \quad \alpha_i = \int_{-1}^{1} \ell_i(x)dx.$$

Ejemplo 8.5.3

Obtener la cuadratura gaussiana de tres nodos. Aproximar mediante ella el valor de la integral $\int_0^1 \operatorname{sen}(x^2)dx$ (utilizar cuatro decimales en las aproximaciones).

En primer lugar, debe determinarse el valor del polinomio de Legendre $P_3(x)$, lo cual se hará sabiendo que $P_0(x) = 1$, $P_1(x) = x$, y de la relación de recurrencia $(n+1)P_{n+1}(x) - (2n+1)xP_n(x) + nP_{n-1}(x) = 0$, $n \geq 1$, a partir de la cual es fácil obtener $P_2(x) = \frac{3x^2-1}{2}$, $P_3(x) = \frac{5x^3-3x}{2}$.

Los nodos serán las raíces de la ecuación $P_3(x) = 0$, esto es $x_1 = -\sqrt{\frac{3}{5}}$, $x_2 = 0$, $x_3 = \sqrt{\frac{3}{5}}$. A partir de estos valores se tiene que:

$$\ell_1(x) = \frac{x\left(x - \sqrt{\frac{3}{5}}\right)}{\left(-\sqrt{\frac{3}{5}}\right)\left(-2\sqrt{\frac{3}{5}}\right)} = \frac{5}{6}x\left(x - \sqrt{\frac{3}{5}}\right)$$

$$\ell_2(x) = \frac{\left(x - \sqrt{\frac{3}{5}}\right)\left(x - \sqrt{\frac{3}{5}}\right)}{\left(\sqrt{\frac{3}{5}}\right)\left(-\sqrt{\frac{3}{5}}\right)} = \frac{5}{3}\left(x^2 - \frac{3}{5}\right)$$

$$\ell_3(x) = \frac{\left(x + \sqrt{\frac{3}{5}}\right)x}{\left(2\sqrt{\frac{3}{5}}\right)(\left(\sqrt{\frac{3}{5}}\right)} = \frac{5}{6}x\left(x - \sqrt{\frac{3}{5}}\right),$$

Integrando, se obtiene:

$$\alpha_1 = \int_{-1}^{1}\left[\frac{5}{6}x\left(x - \sqrt{\frac{3}{5}}\right)\right]dx = \frac{5}{9}$$

$$\alpha_2 = \int_{-1}^{1} \left[\frac{5}{3}\left(x^2 - \frac{3}{5}\right)\right] = \frac{8}{9}$$

$$\alpha_3 = \int_{-1}^{1} \left[\frac{5}{3}x\left(x - \sqrt{\frac{3}{5}}\right)\right] = \frac{5}{9},$$

y por tanto, se tiene la fórmula de cuadratura gaussiana de tres nodos:

$$\int_a^b f(x)dx = \frac{b-a}{2}\left[\frac{5}{9}f\left(\frac{a+b}{2} - \frac{b-a}{2}\sqrt{\frac{3}{5}}\right) + \frac{8}{9}f\left(\frac{a+b}{2}\right) + \right.$$
$$\left. + \frac{5}{9}f\left(\frac{a+b}{2} + \frac{b-a}{2}\sqrt{\frac{3}{5}}\right)\right].$$

Aplicando la fórmula anterior se tiene que

$$\int_0^1 \operatorname{sen} x^2 dx \approx \frac{1}{2}\left[\frac{5}{9}\operatorname{sen}\left(\frac{1}{2} - \frac{1}{2}\sqrt{\frac{3}{5}}\right)^2 + \frac{8}{9}\operatorname{sen}\left(\frac{1}{2}\right)^2 + \operatorname{sen}\left(\frac{1}{2} + \frac{1}{2}\sqrt{\frac{3}{5}}\right)^2\right],$$

esto es

$$\int_0^1 \operatorname{sen} x^2 dx \approx \frac{1}{2}\left[\frac{5}{9}\operatorname{sen} 0.0127 + \frac{8}{9}\operatorname{sen} .2500 + \frac{5}{9}\operatorname{sen} 0.7873\right] = 0.3103\,.$$

Nótese que el valor verdadero de la integral es 0.3103, por tanto mediante el uso de una cuadratura gaussiana de tres nodos se alcanza el verdadero con precisión de 10^{-4}.

♦

Tema 9

Integración de ecuaciones diferenciales

9.1. Introducción

En el presente capítulo se aborda el estudio del problema de la integración de ecuaciones diferenciales ordinarias, tanto a partir de condiciones iniciales como en el caso de condiciones de contorno.

El problema más sencillo es el de condición inicial, el cual se plantea en su versión más simple en los siguientes términos:

obtener una función $y(x)$ de modo que

$$\begin{cases} y'(x) = f(x, y), \; x \in [a, b] \\ y(a) = y_0 \end{cases}$$

Este problema puede ser resuelto bajo ciertas condiciones de modo analítico, si bien en la mayor parte de los casos no es posible obtener una solución del mismo en una forma cerrada.

Los métodos numéricos para resolver este problema dividen el intervalo $[a, b]$ en un determinado número de subintervalos $[x_{i-1}, x_i]$, $i = 1, \ldots, n$ de modo que $a = x_0 < x_1 < \cdots < x_n = b$. Estos métodos se clasifican en métodos de pasos libres y métodos de pasos ligados. Los métodos de pasos libres obtienen aproximaciones a $y(x_{i+1})$ a partir de un único valor $y(x_i)$. Los métodos de pasos ligados obtienen sus aproximaciones a $y(x_{i+k})$ a partir de los valores $y(x_i), y(x_{i+1}), \ldots, y(x_{i+k-1})$.

La solución en un punto cualquiera del intervalo $[a, b]$ puede ser obtenida a partir de la tabla $\{x_i, y(x_i))\}_{i=0}^{n}$ mediante un adecuado proceso de reconstrucción como puede ser la interpolación.

Los llamados problemas de contorno se abordarán de modo somero dado que por limitaciones de este curso no es posible profundizar en ellos. A este fin se esbozarán los métodos de tiro y los métodos de diferencias.

9.2. Métodos de pasos libres

Sea el siguiente problema: dada la ecuación diferencial $y'(x) = f(x, y),\ x \in [a, b]$ con $y(a) = y_0$, encontrar una función $y \in C^1([a, b])$ de modo que se satisfaga la ecuación y la condición inicial. Los métodos de pasos libres pretenden obtener una aproximación a $y(a + h)$ mediante una función $Y(a, h, y_a, f)$.

Entre los métodos de pasos libres más utilizados están los métodos de Taylor, y sobre todo los métodos de Runge-Kutta.

9.2.1. Métodos de Taylor

Los métodos de Taylor están basados en aproximar la solución de la ecuación diferencial mediante el desarrollo en serie de Taylor de la función $y(x)$. Dentro de los métodos de Taylor podemos distinguir los métodos explícitos, los cuales proporcionan directamente $y(x + h)$ y los métodos implícitos, los cuales proporcionan una ecuación que debe satisfacer $y(x + h)$.

Si f es suficientemente derivable, se tiene que

$$y(x+h) = y(x)+y'(x)h+\frac{y''(x)}{2}h^2+\cdots+\frac{y^{k)}(x)}{k!}h^k+\frac{y^{k+1)}(x+\theta h)}{(k+1)!}h^{k+1},\ \theta \in]0,1[,$$

donde las derivadas $y^{k)}(x)$ pueden ser calculadas a partir del esquema

$$\begin{aligned} y'(x) &= f(x, y) \\ y''(x) &= \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx} = \frac{\partial f}{\partial x} + f\frac{\partial f}{\partial y}. \end{aligned}$$

Introduciendo el operador

$$D = \frac{\partial}{\partial x} + f\frac{\partial}{\partial y},$$

y la notación simbólica

$$D^n = \left(\frac{\partial f}{\partial x} + f\frac{\partial f}{\partial y}\right)^n,$$

se tiene que $y^{k+1)}(x) = D^k f$.

9.2.2. Métodos de Taylor explícitos

Los métodos de Taylor explícitos de orden k consisten en desarrollar la función $y(x)$ alrededor del punto x_i para obtener, a partir de este desarrollo, el valor de $y(x_i + h)$ mediante una serie de Taylor truncada en orden $k + 1$. Así, para una ecuación diferencial $y' = f(x, y)$ con condición inicial $y(x_0) = y_0$, la fórmula de integración

$$y_{i+1} = y_i + y_i'h + \frac{1}{2}y_i''h^3 + \cdots + \frac{1}{k!}y_i^{k)}h^k, \quad y_i^{k)} = \left.\frac{d^k y}{dx^k}\right|_{x=x_i}.$$

Ejemplo 9.2.1
Integrar en el intervalo $[0,1]$ mediante un método de Taylor explícito de orden 3 la ecuación diferencial $y' = y^2 \operatorname{sen} x e^{-x}$, $y(0) = 1$. Efectuar la integración mediante un paso $h = 1$, mediante un paso $h = 0.5$ y mediante un paso $h = 0.25$.

En primer lugar, debe obtenerse la fórmula de cuadratura; así, aplicando sobre $y^{k)}$ el operador $D = \frac{\partial}{\partial x} + (y^2 \operatorname{sen} x \exp -x) \frac{\partial}{\partial y}$ se obtiene

$$\begin{aligned} y' &= y^2 \operatorname{sen} x e^{-x} \\ y'' &= 2e^{-2x} y^3 \operatorname{sen}^2 x + e^{-x} y^2 \cos x - e^{-x} y^2 \operatorname{sen} x \\ y''' &= -4e^{-2x} y^3 \operatorname{sen}^2 x + 4e^{-2x} y^3 \cos x \operatorname{sen} x - 2e^{-x} y^2 \cos x + \\ &\quad + e^{-x} \operatorname{sen} x \left(6e^{-2x} y^2 \operatorname{sen}^2 x - 2e^{-x} y \operatorname{sen} x + 2e^{-x} y \cos x\right) y^2 \end{aligned}$$

$$\begin{aligned} y_{i+1} = y_i + y_i^2 \operatorname{sen} x_i e^{-x_i} h + \frac{1}{2} y_i^2 \left[2e^{-2x_i} y_1 \operatorname{sen}^2 x_i + e^{-x_i} \cos x_i - e^{-x_i} \operatorname{sen} x_i\right] h^2 + \\ + \frac{1}{6} \left[-4e^{-2x_i} y_i^3 \operatorname{sen}^2 x_i + 4e^{-2x_i} y_i^3 \cos x_i \operatorname{sen} x_i - 2e^{-x_i} y_i^2 \cos x_i + \right. \\ \left. + e^{-x_i} \operatorname{sen} x_i \left(6e^{-2x_i} y_i^2 \operatorname{sen}^2 x_i - 2e^{-x_i} y_i \operatorname{sen} x_i + 2e^{-x_i} y_i \cos x_i\right) y_i^2\right] h^3 \end{aligned}$$

1. Paso $h = 1.0$ $x_0 = 0$, $x_1 = 1$. Aplicando la fórmula anterior se obtiene $y_1 = 1.16667$.
2. Paso $h = 0.5$ $x_0 = 0$, $x_1 = 0.5$, $x_2 = 1$. Aplicando la fórmula anterior se obtiene en primer lugar $y_1 = 1.08333$, y volviendo a aplicar $y_2 = 1.30563$
3. Paso $h = 0.25$. En este caso se tiene

i	0	1	2	3	4
x_i	0.00	.25	.50	.75	1.00
y_i	1.0000	1.0260	1.0956	1.1975	1.3238

El verdadero valor con cuatro cifras es $y = 1.3259$, el cual se alcanza con un paso de $h = 0.1$. ♦

9.2.3. Métodos de Taylor implícitos

Los métodos de Taylor implícitos de orden k consisten en desarrollar la función $y(x)$ alrededor del punto x_{i+1} para obtener a partir de este desarrollo una ecuación, la cual debe ser resuelta para poder obtener a partir de ella el valor de $y(x_i + h)$. La ecuación se obtiene a partir de una serie de Taylor truncada en orden $k+1$. En este caso a partir del desarrollo:

$$y_i = y_{i+1} - y'_{i+1} h + \frac{1}{2} y''_{i+1} h^2 - \frac{1}{6} y'''_{i+1} h^3 + \cdots + (-1)^k \frac{1}{k!} y^{k)}_{i+1} h^k,$$

se obtiene la fórmula de integración

$$y_{i+1} = y_i + y'_{i+1} h - \frac{1}{2} y''_{i+1} h^2 + \frac{1}{6} y'''_{i+1} h^3 + \cdots + (-1)^{k+1} \frac{1}{k!} y^{k)}_{i+1} h^k,$$

lo que dados los valores de $x_i,\ y_i,\ h$ proporciona una ecuación en y_{i+1} de la forma $y_{i+1} = \Psi(x_i, y_i, h, y_{i+1})$, la cual se debe resolver. Un procedimiento para resolver esta ecuación consiste en tomar una aproximación inicial dada por una fórmula explícita, para a continuación utilizar la ecuación del método implícito de modo iterativo hasta alcanzar la solución de ésta con suficiente precisión; caso de no disponer de tal fórmula y ser el paso suficientemente pequeño, se puede tomar y_i como aproximación inicial.

Ejemplo 9.2.2
Integrar con paso $h = 1$ la ecuación del ejemplo anterior.

La fórmula de cuadratura resultante es

$$y_{i+1} = y_i + \frac{1}{2} y_{i+1}^2 \left[2e^{-2x_{i+1}} y_1 \operatorname{sen}^2 x_{i+1} - e^{-x_{i+1}} \cos x_{i+1} - e^{-x_{i+1}} \operatorname{sen} x_{i+1}\right] h^2 +$$

$$+\frac{1}{6}\left[-4e^{-2x_{i+1}} y_{i+1}^3 \operatorname{sen}^2 x_{i+1} + 4e^{-2x_{i+1}} y_{i+1}^3 \cos x_{i+1} \operatorname{sen} x_{i+1} - 2e^{-x_{i+1}} y_{i+1}^2 \cos x_{i+1} +\right.$$

$$\left.+e^{-x_{i+1}} \operatorname{sen} x_{i+1} \left(6e^{-2x_{i+1}} y_{i+1}^2 \operatorname{sen}^2 x_{i+1} - 2e^{-x_{i+1}} y_{i+1} \operatorname{sen} x_{i+1} + 2e^{-x_{i+1}} y_{i+1} \cos x_{i+1}\right) y_{i+1}^2\right] h^3$$

Teniendo en cuenta que $y_0 = 1$, $x_1 = 1$, y tomando como primera iteración a y_1 el valor $y_{1,0} = 1$, se tiene

n	1	2	3	4	5	6	7	8	9
$y_{1,n}$	1.1982	1.2661	1.2910	1.3002	1.3037	1.3051	1.3055	1.3059	1.3059

Por tanto, $y_1 \approx 1.3059$, valor mucho mas próximo a la solución real que el proporcionado por el método explícito ♦

9.2.4. Métodos de Euler

A continuación se exponen dos métodos de Taylor de primer orden, uno explícito y otro implícito llamados métodos de Euler.

Método de Euler explícito

El método de Taylor explícito de primer orden también recibe el nombre de método de Euler explícito. Para el problema $y' = f(x, y)$, $x \in [a, b]$, $y(a) = y_0$ resulta para el mallado $x_0 = a$, $h = \frac{b-a}{n}$, $x_k = x_0 + k\,h$, $k = 0, \ldots, n$,

$$y_{i+1} = y_i + f(x_i, y_i)h.$$

El error de truncamiento local de este método viene dado por $\frac{y''(\xi)}{2}h^2$, donde $\xi \in [x_i, x_{i+i}]$. Es por tanto un método exacto para polinomios hasta grado uno.

Ejemplo 9.2.3
Integrar la ecuación diferencial $y' = x + y$, entre 0 y 1 sabiendo que $y(0) = 1.$. Utilícese el método explícito de Euler con paso $h = 0.5$ (utilizar dos cifras decimales).

En este caso se tiene que $x_0 = 0.0$, $x_1 = 0.5$, $x_2 = 1.0$. Así:

$$y_1 = y_0 + h\, f(x_0, y_0) = 1.00 + 0.50(0.00 + 1.00) = 1.50$$
$$y_2 = y_1 + h\, f(x_1, y_1) = 1.50 + 0.50(0.50 + 1.50) = 2.50$$

La solución proporcionada por el método de Euler explícito es $y(1) \approx 2.5$.

Nótese que en este caso la solución exacta es conocida y viene dada por $y = 2e^x - x - 1$ y por tanto $y(1) = 2e - 2 = 3.4366$. ♦

Método de Euler implícito

El método de Taylor implícito de primer orden también recibe el nombre de método de Euler implícito. Para el problema $y' = f(x, y)$, $x \in [a, b]$, $y(a) = y_0$ resulta para el mallado $x_0 = a$, $h = \frac{b-a}{n}$, $x_k = x_0 + k\, h$, $k = 0, \ldots, n$

$$y_{i+1} = y_i + f(x_{i+1}, y_{i+1})h.$$

Ejemplo 9.2.4
Integrar la ecuación diferencial $y' = x + y$, entre 0 y 1 sabiendo que $y(0) = 1.0$. Utilícese el método implícito de Euler con paso $h = 0.5$ (efectuar los cálculos con dos cifras decimales). ♦

En este caso se tiene que $x_0 = 0.0$, $x_1 = 0.5$, $x_2 = 1.0$. Así:

$$y_1 = y_0 + 0.5f(0.5, y_1)$$

Tomando como primera aproximación la dada por el método de Euler explícito se tiene

$$y_{1,0} = y_0 + hf(x_0, y_0) = 1.00 + .50(0.00 + 1.00) = 1.50,$$

de donde aplicando la fórmula iterativa

$$y_{1,k+1} = y_0 + 0.5f(0.5, y_{1,k}),$$

se tiene

k	1	2	3	4	5	6	7	8
$y_{1,k}$	2.00	2.25	2.38	2.47	2.48	2.49	2.50	2.50

Para y_2 se tiene, en primer lugar, la aproximación inicial dada por el método de Euler explícito

$$y_{2,0} = 2.50 + 0.50f(0.50, 2.50) = 4.25,$$

a partir de la cual se tiene

k	1	2	3	4	5	6	7	8
$y_{2,k}$	5.12	5.56	5.78	5.89	5.95	5.97	5.99	5.99

por tanto el método de Euler implícito nos proporciona la solución $y(1) \approx 5.99$.

θ-métodos

Los métodos de Euler explícito e implícito pueden ser incluidos en los llamados θ-métodos. Sea $\theta \in [0,1]$, un θ-método es una combinación lineal convexa de un método explícito y de un método implícito según los valores de un parámetro θ. Así, se tiene

$$y_{i+1} = y_i + [\theta f(x_i, y_i) + (1-\theta) f(x_{i+1}, y_{i+1})]\, h.$$

El error de truncamiento de estos métodos viene dado por

$$E = \left[\frac{\theta}{2} y''(\xi_1) - \frac{(1-\theta)}{2} y''(\xi_2)\right] h^2, \xi_1, \xi_2 \in]x_i, x_{i+1}].$$

Por tanto, $\xi_2 = \xi_1 + t\,h,\ t \in]-1, 1[$, a partir de lo cual

$$y''(\xi_2) = y''(\xi_1 + t\,h) = y''(\xi_1) + y'''(\xi_1 + \alpha t\,h)h,\ \alpha \in]0,1[,$$

de donde

$$E = \frac{2\theta - 1}{2} y''(\xi_1)h^2 - \frac{1-\theta}{2} y'''(\xi_1 + \alpha t\,h)h^3,$$

el cual es de orden $O(h^2)$, excepto en el caso de $\theta = \frac{1}{2}$ en el cual $O(h^3)$. En este último caso el método es exacto para polinomios hasta grado dos. El θ-método resulta para $\theta = \frac{1}{2}$:

$$y_{i+1} = y_i + \frac{1}{2}\left[f(x_i, y_i) + f(x_{i+1}, y_{i+1})\right] h.$$

Ejemplo 9.2.5

Integrar la ecuación diferencial $y' = x + y$, entre 0 y 1 sabiendo que $y(0) = 1.0$. Utilícese un θ-métodos con $\theta = 0.2$ y paso $h = 0.5$.

Para $\theta = 0.2$, se tiene:

$$y_1 = 1.0000 + 0.5\ [0.2\, f(0.00, 1.0000) + 0.8\, f(0.50, y_1)]\,,$$

lo que resulta una ecuación implícita en y_1. Tomando como aproximación inicial la dada por el método de Euler explícito, $y_{1,0} = 1.0 + 0.5\, f(0.0, 1.0) = 1.5$, se tiene:

	n	1	2	3	4	5	6	7	8
$y_{1,n}$	1.500	1.900	2.060	2.124	2.150	2.160	2.164	2.166	2.166

a continuación, para y_2, se tiene:

$$y_1 = 2.166 + 0.5\ [0.2\, f(0.50, 2.166) + 0.8\, f(1.0, y_2)]\,,$$

ecuación implícita en y_2. Tomando como aproxiamción inicial $y_2 = y_1 + 0.5 f(0.5, y_1)$, se tiene:

	n	1	2	3	4	5	6	7	8	9
$y_{2,n}$	1.5	3.433	4.206	4.515	4.639	4.688	4.708	4.716	4.719	4.720

Por tanto $y(1) = 4.720$.

♦

Método del punto medio

Otra variante del método de Euler es el llamado método del punto medio, el cual consiste en desarrollar tanto y_i como y_{i+1} alrededor del punto $x_i + \frac{h}{2}$. Denotando por $x_{i+\frac{1}{2}} = x_i + \frac{h}{2}$, y por $y_{i+\frac{1}{2}} = y(x_{i+\frac{1}{2}})$ se tiene

$$y_{i+1} = y_{i+\frac{1}{2}} + f(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}})\frac{h}{2} + \frac{1}{2}y''(\xi_2)\frac{h^2}{4}$$
$$y_i = y_{i+\frac{1}{2}} - f(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}})\frac{h}{2} + \frac{1}{2}y''(\xi_1)\frac{h^2}{4},$$

restando ambas expresiones, se tiene

$$y_{i+1} - y_i = f(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}})h + [y''(\xi_2) - y''(\xi_1)]\,\frac{h^3}{8},$$

lo cual proporciona el llamado método del punto medio

$$y_{i+1} - y_i = f(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}})h.$$

El valor de $y_{i+\frac{1}{2}}$ puede ser aproximado mediante $y_{i+\frac{1}{2}} = y_i + \frac{h}{2}f(x_0, y_0)$.

El orden de dicho método viene dado por

$$E = y''(\xi_1)\frac{h^2}{8} - y''(\xi_2)\frac{h^2}{8} = -y'''(\xi)y''(\xi_1)\frac{h^2}{8}, \quad \xi \in]m, M[,$$

donde $m = \text{mín}\{\xi_1, \xi 2\}$, $M = \text{máx}\{\xi_1, \xi_2\}$.

Puede observarse que el método es de segundo orden, siendo en este caso la constante multiplicativa menor que para el θ-método con $\theta = \frac{1}{2}$.

Ejemplo 9.2.6
Integrar la ecuación diferencial $y' = x + y$, entre 0 y 1 sabiendo que $y(0)1.0$. Utilícese el método del punto medio con paso $h = 0.5$.

En este caso se tiene que $x_0 = 0.0$, $x_1 = 0.5$, $x_2 = 1.0$; así,

$$y_1 = 1.0000 + .5000f[0.0000 + 0.2500, 1.0000 + 0.2500f(0.0000, 1.0000)] = 1.7500$$
$$y_2 = 1.7500 + .5000f[0.5000 + 0.2500, 1.7500 + 0.2500f(.50000, 1.7500)] = 3.2812$$

♦

9.2.5. Métodos de Runge-Kutta

Los métodos de Taylor presentan el inconveniente de tener la necesidad de evaluar derivadas de la función, lo cual al estar ésta dada de forma implícita puede conducir a elevar el orden de las fórmulas a desarrollos muy complicados. Los métodos de Euler son sencillos de implementar, pero presentan el inconveniente de requerir un muy elevado número de pasos aún cuando la precisión requerida en la solución no sea excesivamente elevada.

Los θ-métodos utilizan la evaluación de la derivada en varios puntos, consiguiéndose en algunos casos ($\theta = \frac{1}{2}$) elevar el orden del método.

La idea principal de los métodos de Runge-Kutta consiste en buscar aproximaciones a la solución en puntos intermedios del intervalo $[x_i, x_{i+1}]$ y utilizar una combinación lineal de los valores de la derivada en varias de estas aproximaciones para obtener un valor de y_{i+1} exacto hasta orden p en h.

Sea la ecuación diferencial $y' = f(x, y),\ y(x_0) = y_0$ y sea $x_i = x_0 + i\,h$ con $h > 0$. Para evaluar $y(x_{i+1})$ una vez conocido el valor de y_i se procede como sigue: sean $0 \leq \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_r \leq 1$ y sea $\gamma_1, \ldots, \gamma_r \in [0, 1]$ de modo que $\sum\limits_{i=1}^{r} \gamma_i = 1$. Los métodos de Runge-Kutta pretenden evaluar y_{i+1} como

$$y_{i+1} = y_i + \sum_{i=1}^{r} \gamma_i k_i,$$

donde

$$k_i = hf(x_i + \alpha_i\,h, y_i + \sum_{j=1}^{r} \beta_{i,j} k_j), \quad \sum_{j=1}^{r} \beta_{i,j} = \alpha_i.$$

Los coeficientes γ_i, α_i, $\beta_{i,j}$, $i, j = 1, \ldots, r$ deben determinarse de modo que el desarrollo en serie de Taylor de la aproximación $y_i + \sum_{i=1}^{r} \gamma_i k_i$ coincida con el desarrollo de $y(x_i + h)$ hasta orden p en h.

Los métodos de Runge-Kutta se clasifican en explícitos, cuando los valores de k_i pueden ser evaluados en función de $k_1, k_2, \ldots, k_{i-1}$, e implícitos, cuando lo anterior no es posible. En los métodos explícitos, la matriz $(\beta_{i,j})$ satisface $\beta_{i,j} = 0,\ \forall i \leq j$. En los métodos implícitos debe resolverse en cada paso un sistema de ecuaciones de la forma

$$\begin{array}{rcl} k_1 &=& f(x_i + \alpha_1\,h, y_i + \beta_{1,1}k_1 + \beta_{1,2}k_2 + \cdots + \beta_{1,p}k_p) \\ k_2 &=& f(x_i + \alpha_2\,h, y_i + \beta_{2,1}k_1 + \beta_{2,2}k_2 + \cdots + \beta_{2,p}k_p) \\ \cdots &=& \cdots \\ k_p &=& f(x_i + \alpha_p\,h, y_i + \beta_{p,1}k_1 + \beta_{p,2}k_2 + \cdots + \beta_{p,p}k_p). \end{array}$$

El proceso de obtención de los coeficientes es costoso, por ello, dado que el procedimiento es el mismo, se expondrá únicamente los métodos de Runge-Kutta de orden dos con dos evaluaciones de la función.

Métodos de Runge-Kutta de orden dos con dos evaluaciones de la función

Sea el problema de condición inicial para la ecuación diferencial $y' = f(x, y)$ con $y(x_0) = y_0$ donde $f(x, y)$ es una función de clase $C^3([a, b])$ de modo que $x_0 \in [a, b]$, $x_1 = x_0 + h \in [a, b]$, $h > 0$. En estas condiciones se tiene que

$$y_1 = y_0 + \gamma_1 k_1 + \gamma_2 k_2,$$

donde

$$k_1 = h\,f(x_0 + \alpha_1 h, y_0 + \beta_{1,1}k_1 + \beta_{1,2}k_2)$$
$$k_2 = h\,f(x_0 + \alpha_2 h, y_0 + \beta_{2,1}k_1 + \beta_{2,2}k_2).$$

Desarrollando $y(x_0+h)$ hasta orden dos en h se tiene

$$y(x_0+h) = y(x_0) + y'(x_0)h + \frac{1}{2}y''(x_0)h^2 + O(h^3),$$

esto es

$$y(x_0+h) = y(x_0) + y'(x_0)h + \frac{1}{2}\left.\frac{\partial f}{\partial x}\right|_{(x_0,y_0)} h^2 + \frac{1}{2}f(x_0,y_0)\left.\frac{\partial f}{\partial y}\right|_{(x_0,y_0)} h^2 + O(h^3).$$

Para desarrollar y_1 hasta orden dos, y puesto que k_1, k_2 son de primer orden en h, bastará desarrollar $f(x+\Delta x, y+\Delta y)$ hasta orden uno en Δx, Δy. Así se tiene

$$f(x+\Delta x, y+\Delta y) = f(x,y) + \left.\frac{\partial f}{\partial x}\right|_{(x,y)} \Delta x + \left.\frac{\partial f}{\partial y}\right|_{(x,y)} \Delta y + O(h^2)$$

Por tanto

$$\begin{aligned}
k_1 &= f(x_0,y_0)h + \left.\frac{\partial f}{\partial x}\right|_{(x_0,y_0)} \alpha_1 h^2 + \left.\frac{\partial f}{\partial y}\right|_{(x_0,y_0)} \left(\beta_{1,1}k_1 + \beta_{1,2}k_2\right) h + O(h^3) \\
k_2 &= f(x_0,y_0)h + \left.\frac{\partial f}{\partial x}\right|_{(x_0,y_0)} \alpha_2 h^2 + \left.\frac{\partial f}{\partial y}\right|_{(x_0,y_0)} \left(\beta_{2,1}k_1 + \beta_{2,2}k_2\right) h + O(h^3).
\end{aligned}$$

En el segundo miembro de estas igualdades aparece k_1, k_2 en términos de orden uno en h, por tanto, para obtener los desarrollos en segundo orden bastará aproximar en orden uno k_1, k_2, esto es $k_1 = k_2 = hf(x_0,y_0)$ de donde se tiene

$$\begin{aligned}
k_1 &= f(x_0,y_0)h + \left.\frac{\partial f}{\partial x}\right|_{(x_0,y_0)} \alpha_1 h^2 + f(x_0,y_0)\left.\frac{\partial f}{\partial y}\right|_{(x_0,y_0)} \left(\beta_{1,1} + \beta_{1,2}\right) h^2 + O(h^3) \\
k_2 &= f(x_0,y_0)h + \left.\frac{\partial f}{\partial x}\right|_{(x_0,y_0)} \alpha_2 h^2 + f(x_0,y_0)\left.\frac{\partial f}{\partial y}\right|_{(x_0,y_0)} \left(\beta_{2,1} + \beta_{2,2}\right) h^2 + O(h^3).
\end{aligned}$$

Entonces, para y_1 se tiene

$$\begin{aligned}
y_1 = y_0 &+ (\gamma_1+\gamma_2)f(x_0,y_0)h + (\gamma_1\alpha_1 + \gamma_2\alpha_2)\left.\frac{\partial f}{\partial x}\right|_{(x_0,y_0)} \alpha_1 h^2 + \\
&+ \left[\gamma_1\left(\beta_{1,1}+\beta_{1,2}\right) + \gamma 2\left(\beta_{2,1}+\beta_{2,2}\right)\right] f(x_0,y_0)\left.\frac{\partial f}{\partial y}\right|_{(x_0,y_0)} h^2 + O(h^3).
\end{aligned}$$

Por tanto, deben satisfacerse las igualdades

$$\begin{aligned}
&\gamma_1 + \gamma_2 = 1 \\
&\gamma_1\alpha_1 + \gamma_2\alpha_2 = \frac{1}{2} \\
&\gamma_1\left(\beta_{1,1}+\beta_{1,2}\right) + \gamma_2\left(\beta_{2,1}+\beta_{2,2}\right) = \frac{1}{2}.
\end{aligned}$$

Por razones de compatibilidad del sistema, y dado que la evaluación de la función $f(x, y)$ debe realizarse en un punto lo más próximo posible a $f(x, y(x))$, se tiene que $\beta_{1,1} + \beta_{1,2} = \alpha_1$, $\beta_{2,1} + \beta_{2,2} = \alpha_2$. Por tanto, debe verificarse

$$\begin{aligned}
&\gamma_1 + \gamma_2 = 1 \\
&\gamma_1\alpha_1 + \gamma_2\alpha_2 = \frac{1}{2} \\
&\beta_{1,1} + \beta_{1,2} = \alpha_1 \\
&\beta_{2,1} + \beta_{2,2} = \alpha_2.
\end{aligned}$$

Ejemplo 9.2.7
Encontrar un método de Runge-Kutta explícito de segundo orden con dos evaluaciones de la función en los puntos x_0, x_1.

Con las condiciones del enunciado, se tiene que $\alpha_1 = 0$, $\alpha_2 = 1$, $\beta_{1,1} = \beta_{1,2} = \beta_{2,2} = 0$, por tanto, $\beta_{1,2} = 1$, $\gamma_1 = \frac{1}{2}$, $\gamma_2 = \frac{1}{2}$. Así, resulta

$$\begin{aligned}
k_1 &= hf(x_0, y_0) \\
k_2 &= hf(x_0 + h, y_0 + k_1) \\
y_1 &= y_0 + \frac{1}{2}k_1 + \frac{1}{2}k_2.
\end{aligned}$$

Este método explícito de Runge-Kutta de segundo orden se conoce como método de Heun. ♦

Ejemplo 9.2.8
Obtener un método de Runge-Kutta explícito de segundo orden mediante dos evaluaciones de la función en los puntos x_0, $x_0 + \frac{1}{2}h$.

En este caso resulta $\alpha_1 = 0$, $\alpha_1 = \frac{1}{2}$, $\beta_{1,1} = \beta_{1,1} = \beta_{2,2} = 0$, $\beta_{2,1} = \frac{1}{2}$, $\gamma_1 = 0$, $\gamma_2 = 1$, por tanto se tiene

$$\begin{aligned}
k_1 &= hf(x_0, y_0) \\
k_2 &= hf(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}k_1) \\
y_1 &= y_0 + k_2.
\end{aligned}$$

Este método explícito de Runge-Kutta de segundo orden se conoce como método del punto medio modificado. ♦

Métodos de Runge-Kutta de orden superior

Los métodos de Runge-Kutta de órdenes mayores se obtienen de modo similar solo que con mucho más esfuerzo, por esta razón se omite su deducción. A continuación se expone el método de Runge–Kutta clásico de cuarto orden, dado que es el más utilizado.

Método de Runge-Kutta clásico de cuarto orden

Este método utiliza la evaluación de la función en los puntos x_i, $x_i + \frac{h}{2}$ y $x_i + h$. Así, se tiene

$$\begin{aligned}
k_1 &= h\,f\,(x_i, y_i)\\
k_2 &= h\,f\left(x_i + \frac{h}{2}, y_0 + \frac{k_1}{2}\right)\\
k_3 &= h\,f\left(x_i + \frac{h}{2}, y_0 + \frac{k_2}{2}\right)\\
k_2 &= h\,f\,(x_i + h, y_0 + k_3)\\
y_{i+1} &= y_i + \frac{1}{6}\left(k_1 + 2\,k_2 + 2\,k_3 + k_4\right).
\end{aligned}$$

Ejemplo 9.2.9
Integrar la ecuación diferencial $y' = x + y$, entre 0 y 1 sabiendo que $y(0) = 1.0$. Utilícese el método de Runge–Kutta clásico de cuarto orden con paso $h = .5$.

En este caso se tiene que $x_0 = 0$, $x_1 = 0.5$, $x_2 = 1.0$; así:

x_i	y_i	k_1	k_2	k_3	k_4	$y_i + 1$
0.0	1.0000	0.5000	0.7500	0.8125	1.1563	1.7969
0.5	1.7969	1.1484	1.5606	1.6636	2.2302	3.4347

♦

9.2.6. Métodos de pasos ligados

Los métodos de pasos ligados son métodos en los cuales se requiere el conocimiento del valor de la función sobre un número k de nodos $x_i, x_{i+1}, \ldots, x_{i+k-1}$ para poder determinar el valor de la función en el nodo y_{i+k}, lo que implica una relación funcional de la forma $y_{i+k} = \Psi(x_i, h, y_i, y_{i+1}, \ldots, y_{i+k}, f)$. En el estudio que aquí se presenta únicamente se abordarán los llamados métodos lineales multipaso, los cuales son métodos de la forma

$$a_0 y_i + a_1 y_{i+1} + \cdots + a_{k-1} y_{i+k-1} + a_k y_{i+k} = h(b_0 f_i + \cdots + b_{k-1} f_{i+k-1} + b_k f_{i+k},$$

donde $a_k \neq 0$, $|a_0| + |b_0| \neq 0$, $f_j = f(x_j, y_j)$. Por conveniencia, en lo sucesivo se tomará $a_k = 1$.

Un método lineal multipaso se dice explícito cuando $b_k = 0$ e implícito cuando $b_k \neq 0$. Un método lineal multipaso se dice de orden n cuando la fórmula del método es exacta hasta orden n en h, lo cual implica que los desarrollos en serie de Taylor de ambos miembros coinciden hasta orden n. Para obtener métodos lineales multipasos existen tres procedimientos fundamentales, los desarrollos en serie de Taylor, la interpolación osculatoria y la integración numérica.

Los métodos lineales multipaso son fáciles de programar, pero presentan el inconveniente de necesitar de un método de pasos libres de orden mayor o igual que él para calcular los puntos $y_1, y_2, \ldots, y_{k-1}$ necesarios para iniciar el cálculo.

Métodos lineales multipaso obtenidos mediante desarrollo en serie

Sea un método lineal de k pasos dado por la ecuación

$$a_0 y_i + a_1 y_{i+1} + \cdots + a_{k-1} y_{i+k-1} + a_k y_{i+k} = h(b_0 f_i + \cdots + b_{k-1} f_{i+k-1} + b_k f_{i+k}).$$

La condición que debe cumplir para ser un método de orden n es que los desarrollos en serie de Taylor de ambos miembros coincidan hasta orden n. De este modo, se obtiene un conjunto de ecuaciones para los coeficientes $a_i, \ldots, a_{i+k-1}, b_i, \ldots, b_{i+k}$ las cuales deben ser satisfechas; caso de ser incompatibles no existiría un método lineal multipaso de tales características.

Ejemplo 9.2.10
Encontrar un método lineal de cuatro pasos explícito de orden cuatro.

El método vendrá dado por la relación

$$a_0 y_i + a_1 y_{i+1} + a_2 y_{i+2} + a_3 y_i + 3 + y_{i+4} = h(b_0 f_i + b_1 f_{i+1} + b_2 f_{i+2} + b_3 f_{i+3})$$

Los desarrollos en serie de Taylor y_{i+1}, y_{i+2},y_{i+3} hasta orden cuatro en h vienen dados por

$$\begin{aligned}
y_{i+1} &= y_i + y_i' h + \frac{1}{2} y_i'' h^2 + \frac{1}{6} y_i''' h^3 + \frac{1}{24} y_i^{4)} h^4 + O(h^5) \\
y_{i+2} &= y_i + 2y_i' h + 2y_i'' h^2 + \frac{4}{3} y_i''' h^3 + \frac{2}{3} y_i^{4)} h^4 + O(h^5) \\
y_{i+3} &= y_i + 3y_i' h + \frac{9}{2} y_i'' h^2 + \frac{9}{2} y_i''' h^3 + \frac{27}{8} y_i^{4)} h^4 + O(h^5) \\
y_{i+4} &= y_i + 4y_i' h + 8y_i'' h^2 + \frac{32}{3} y_i''' h^3 + \frac{32}{3} y_i^{4)} h^4 + O(h^5).
\end{aligned}$$

Por otra parte, se tiene que $f_i = y_i'$ y los desarrollos en serie de Taylor f_{i+1}, f_{i+2},f_{i+3} hasta orden tres en h son:

$$\begin{aligned}
f_{i+1} &= y_i' + y_i'' h + \frac{1}{2} y_i''' h^2 + \frac{1}{6} y_i^{4)} h^3 + O(h^4) \\
f_{i+2} &= y_i' + 2y_i'' h + 2y_i''' h^2 + \frac{4}{3} y_i^{4)} h^3 + O(h^4) \\
f_{i+3} &= y_i' + 3y_i'' h + \frac{9}{2} y_i''' h^2 + \frac{9}{2} y_i^{4)} h^3 + O(h^4).
\end{aligned}$$

Por tanto, se tiene, igualando términos del mismo orden en el primer y segundo

miembro del método:

$$\begin{aligned} a_0 + a_1 + a_2 + a_3 + 1 &= 0 \\ a_1 + 2a_2 + 3a_3 + 4 &= b_0 + b_1 + b_2 + b_3 \\ \frac{1}{2}a_1 + \frac{4}{3}a_2 + \frac{2}{3}a_3 + 8 &= b_1 + 2b_2 + 3b_3 \\ \frac{1}{6}a_1 + \frac{2}{3}a_2 + \frac{9}{2}a_3 + \frac{32}{3} &= \frac{1}{2}b_1 + \frac{4}{3}b_2 + \frac{2}{3}b_3 \\ \frac{1}{24}a_1 + \frac{2}{3}a_3 + \frac{27}{8}a_3 + \frac{32}{3} &= \frac{1}{6}b_1 + \frac{2}{3}b_2 + \frac{9}{2}b_3. \end{aligned}$$

En este sistema hay más incógnitas que ecuaciones, por tanto, se puede fijar el valor de algunas de ellas, haciendo $a_0 = a_1 = a_2 = 0$ se tiene que $a_3 = -1$. Así:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 3 \\ 0 & \frac{1}{2} & 2 & \frac{9}{3} \\ 0 & \frac{1}{6} & \frac{4}{3} & \frac{9}{2} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{7}{2} \\ \frac{37}{6} \\ \frac{175}{24} \end{bmatrix},$$

y por tanto, $b_0 = -\frac{9}{24}$, $b_1 = \frac{37}{24}$, $b2 = -\frac{59}{24}$, $b3 = \frac{55}{24}$. Finalmente, la fórmula de integración resulta:

$$y_{i+4} = y_{i+3} + \frac{h}{24}[-9f_i + 37f_{i+1} - 59f_{i+2} + 55f_{i+3}]$$

♦

Ejemplo 9.2.11

Encontrar un método lineal de dos pasos implícito de orden cuatro. El método vendrá dado por la relación

$$a_0 y_i + a_1 y_{i+1} + y_{i+2} = h(b_0 f_i + b_1 f_{i+1} + b_2 f_{i+2}).$$

Desarrollando y_{i+1}, y_{i+2}, y'_{i+1}, y'_{i+2} hasta orden cuatro alrededor de x_i, se tiene

$$\begin{aligned} y_{i+1} &= y_i + \ y'_i h + \frac{1}{2}y''_i h^2 + \frac{1}{6}y'''_i h^3 + \frac{1}{24}y_i^{4)} h^4 + O(h^5) \\ y_{i+2} &= y_i + 2y'_i h + 2y''_i h^2 + \frac{4}{3}y'''_i h^3 + \frac{2}{3}y_i^{4)} h^4 + O(h^5) \\ f_{i+1} &= y'_i + \ y''_i h + \frac{1}{2}y'''_i h^2 + \frac{1}{6}y_i^{4)} h^3 + O(h^4) \\ f_{i+2} &= y'_i + 2y''_i h + 2y'''_i h^2 + \frac{4}{3}y_i^{4)} h^3 + O(h^4), \end{aligned}$$

por tanto, debe cumplirse

$$\begin{aligned} a_0 + a_1 + 1 &= 0 \\ a_1 + 2 &= b_0 + b_1 + b_2 \\ \frac{1}{2}a_1 + 2 &= b_1 + 2b_2 \\ \frac{1}{6}a_1 + \frac{4}{3} &= \frac{1}{2}b_1 + 2b_2 \\ \frac{1}{24}a_1 + \frac{2}{3} &= \frac{1}{6}b_1 + \frac{4}{3}b_2, \end{aligned}$$

lo cual constituye un sistema lineal de cinco ecuaciones con cinco incógnitas cuya solución viene dada por $a_0 = -1$, $a_1 = 0$, $b_0 = \frac{1}{3}$, $b_1 = \frac{4}{3}$, $b_2 = \frac{1}{3}$; por tanto, se tiene la fórmula

$$y_{i+2} = y_1 + \frac{h}{3}[f_i + 4f_{i+1} + f_{i+2}].$$

♦

Métodos lineales multipaso obtenidos mediante integración numérica

Considérese la ecuación diferencial $y' = f(x, y)$ con la condición inicial $y(x_0) = y_0$ y sea un mallado regular $\omega = \{x_i = x_0 + i\,h \,:\, i \in \mathbb{Z}\}$. Supongamos conocida la función $y(x)$ sobre los puntos $x_i, x_{i+1}, \ldots, x_{i+k-1}$ y sean $y_i, \ldots, y_{i+k-1}$ sus respectivos valores. Un método alternativo para obtener métodos lineales multipaso consiste en integrar la ecuación diferencial. Para ello, se procede a aproximar el segundo miembro por un polinomio que interpole la función $f(x, y)$. Si el conjunto de nodos sobre los que se interpola no contiene al nodo x_{i+k}, el método resultante será explícito, en caso contrario resultará implícito.

Métodos explícitos

Sea la ecuación diferencial dada por $y' = f(x, y)$ y sean conocidos los valores de $y_i, y_{i+1}, \ldots, y_{i+k-1}$. Sea x_{i+r} con $0 \leq r < k$ y considérese

$$\int_{x_{i+r}}^{x_{i+k}} y'(x)dx = \int_{x_{i+r}}^{x_{i+k}} f(x, y(x))dx.$$

Para evaluar la integral del segundo miembro se procede a sustituir $f(x, y(x))$ por su polinomio interpolador de Lagrange sobre los nodos $x_1, \ldots, x_{i+k-1}$ obteniéndose

$$y_{i+k} - y_{i+r} = \int_{x_{i+r}}^{x_{i+k}} \sum_{j=0}^{k-1} \ell_j(x) f_j dx = \sum_{j=0}^{k-1} \left[\int_{x_{i+r}}^{x_{i+k}} \ell_j(x)dx \right] f_j,$$

donde $\ell_j(x)$ es la j-ésima función de base de la interpolación de Lagrange y $f_j = f(x_j, y_j)$.

Nótese que en este caso la función $f(x, y(x))$ se interpola en el intervalo $[x_i, x_{i+k-1}]$, por tanto, para efectuar la integración se utilizan valores extrapolados de la función, lo cual implica, en general, errores más altos que cuando se interpola.

Ejemplo 9.2.12

Obtener un método lineal de tres pasos integrando entre x_{i+2} y x_{i+3}. ¿Qué se puede decir del orden del método?

En primer lugar para evaluar las integrales $\int_{x_{i+2}}^{x_{i+3}} \ell_j(x)dx$ se introduce la nueva variable t de modo que $x = x_i + t\,h$. Así, se tiene que $dx = h\,dt$, $t(x_k) = k$, por tanto se tiene

$$\int_{x_{1+2}}^{x_{i+3}} \ell_j(x)dx = h \int_2^3 \ell_j(x(t))dt.$$

Las funciones de base resultan

$$\begin{aligned}
\ell_0(x(t)) &= \frac{(t-1)(t-2)}{(-1)(-2)} = \frac{1}{2}(t-1)(t-2) \\
\ell_1(x(t)) &= \frac{t(t-2)}{1(-1)} = -t(t-2) \\
\ell_2(x(t)) &= \frac{t(t-1)}{2\cdot 1} = \frac{1}{2}t(t-1),
\end{aligned}$$

de donde

$$\begin{aligned}
&\int_{x_{1+2}}^{x_{i+3}} \ell_0(x)dx = h\int_2^3 \ell_0(x(t))dt = h\int_2^3 \frac{1}{2}(t-1)(t-2)dt = \frac{5}{12}h \\
&\int_{x_{1+2}}^{x_{i+3}} \ell_1(x)dx = h\int_2^3 \ell_1(x(t))dt = -h\int_2^3 t(t-2)dt = -\frac{4}{3} \\
&\int_{x_{1+2}}^{x_{i+3}} \ell_2(x)dx = h\int_2^3 \ell_2(x(t))dt = h\int_2^3 \frac{1}{2}t(t-1) = \frac{23}{12},
\end{aligned}$$

por tanto, el método lineal multipaso pedido es

$$y_{i+3} - y_{i+2} = \frac{h}{12}[5f_i - 16f_{i+1} + 23f_{i+2}].$$

Del orden del método se puede decir que si y es un polinomio de grado menor o igual que tres, $f'(x, y(x))$ es de grado menor o igual que dos, por tanto, la interpolación de f' con tres nodos resulta exacta, y por ello la integral. Por tanto, el método tiene un error de truncamiento al menos de tercer orden. ♦

Métodos implícitos

Sea la ecuación diferencial dada por $y' = f(x, y)$ y sean conocidos los valores de $y_i, y_{i+1}, \dots, y_{i+k-1}$. Sea x_{i+r} con $0 \le r < k$ y considérese la igualdad

$$\int_{x_{i+r}}^{x_{i+k}} y'(x)dx = \int_{x_{i+r}}^{x_{i+k}} f(x, y(x))dx.$$

Para evaluar la integral del segundo miembro se procede a sustituir $f(x, y(x))$ por su polinomio interpolador de Lagrange sobre los nodos $x_1, \dots, x_{i+k}$ obteniéndose

$$y_{i+k} - y_{i+r} = \int_{x_{i+r}}^{x_{i+k}} \sum_{j=0}^{k} \ell_j(x) f_j dx = \sum_{j=0}^{k} \left[\int_{x_{i+r}}^{x_{i+k}} \ell_j(x)dx\right] f_j,$$

donde $\ell_j(x)$ es la j-ésima función de base de la interpolación de Lagrange y $f_j = f(x_j, y_j)$.

En este caso no se precisa de extrapolación, por lo que el error debido a la integración se espera, en general, menor.

Ejemplo 9.2.13
Obtener un método lineal de tres pasos integrando entre x_{i+2} y x_{i+3}. ¿Qué se puede decir del orden del método?

Al igual que en el ejemplo anterior se transforman las integrales mediante el cambio de variable $x = x_i + t\,h$. En este caso, las funciones de base resultan

$$\begin{aligned}
\ell_0(x(t)) &= \frac{(t-1)(t-2)(t-3)}{(-1)(-2)(-3)} = -\frac{1}{6}(t-1)(t-2)(t-3) \\
\ell_1(x(t)) &= \frac{t(t-2)(t-3)}{1(-1)(-2)} = \frac{1}{2}t(t-2)(t-3) \\
\ell_2(x(t)) &= \frac{t(t-1)(t-3)}{2\cdot 1(-1)} = -\frac{1}{2}t(t-1)(t-3) \\
\ell_3(x(t)) &= \frac{t(t-1)(t-2)}{3\cdot 2\cdot 1} = \frac{1}{6}t(t-1)(t-2).
\end{aligned}$$

El valor de las integrales viene dado por

$$\begin{aligned}
\int_{x_{1+2}}^{x_{i+3}} \ell_0(x)dx &= h\int_2^3 \ell_0(x(t))dt = -h\int_2^3 \frac{1}{6}(t-1)(t-2)(t-3)dt = \frac{1}{24}h \\
\int_{x_{1+2}}^{x_{i+3}} \ell_1(x)dx &= h\int_2^3 \ell_1(x(t))dt = h\int_2^3 \frac{1}{2}t(t-2)(t-3)dt = -\frac{5}{24} \\
\int_{x_{1+2}}^{x_{i+3}} \ell_2(x)dx &= h\int_2^3 \ell_2(x(t))dt = -h\int_2^3 \frac{1}{2}t(t-1)(t-3) = \frac{19}{24} \\
\int_{x_{1+2}}^{x_{i+3}} \ell_2(x)dx &= h\int_2^3 \ell_2(x(t))dt = h\int_2^3 \frac{1}{2}t(t-1)(t-2) = \frac{3}{8},
\end{aligned}$$

por tanto, el método resulta

$$y_{i+3} - y_{i+2} = \frac{h}{24}[f_i - 5f_{i+1} + 19f_{i+2} + 12f_{i+3}].$$

El orden del método es al menos cuatro, pues si $y(x)$ es un polinomio de grado menor o igual que cuatro, su derivada $f(x, y(x))$ es un polinomio de grado menor o igual que tres, con lo que la aproximación dada por la interpolación sobre los cuatro nodos $x_i, x_{i+1}, x_{i+2}, x_{i+3}$ es exacta. ♦

Métodos lineales multipaso obtenidos mediante interpolación osculatoria

El método de interpolación consiste en encontrar un polinomio $P(x)$ que satisfaga las condiciones $P(x_i) = y_i, \ldots, P(x_{i+k-1}) = y_{k-1}, P'(x_i) = f_i, \ldots, P'(x_{i+k-1}) = f_{i+k-1}$ en el caso de un método explícito. En el caso implícito debe satisfacerse además $P'(x_{i+k}) = f_{i+k}$. El valor de y_{i+k} se obtiene a partir del polinomio interpolante como $y_{i+k} = P(x_{i+k})$.

Ejemplo 9.2.14
Encontrar un método lineal multipaso explícito de dos pasos del mayor orden posible.

Considérese los nodos x_i, x_{i+1}, x_{i+2} igualmente espaciados mediante un paso h. El polinomio de mayor grado que puede interpolar la función y su derivada en los nodos x_i, x_{i+1} es de grado tres, el cual podemos buscar en su forma de Hermite:

$$H_3(x) = A_0(x)y_i + A_1(x)y_{i+1} + B_0(x)f_i + B_1(x)f_{i+1},$$

las funciones de base A_0, A_1, B_0, B_1 vienen dadas por

$$A_0(x) = [1 - 2(x - x_i)\ell_0'(x_i)]\ell_0^2(x), \qquad A_1(x) = [1 - 2(x - x_{i+1})\ell_0'(x_{i+1})]\ell_1^2(x)$$

$$B_0(x) = (x - x_i)\ell_0^2(x), \qquad B_1(x) = (x - x_{i+1})\ell_1^2(x),$$

donde $\ell_j(x)$ son las funciones de base de la interpolación de Lagrange sobre los nodos x_i, x_{i+1}, por tanto:

$$\ell_0(x) = -\frac{x - x_{i+1}}{x_i - x_{i+1}} = -\frac{1}{h}(x - x_{i+1}), \qquad \ell_0'(x_i) = -\frac{1}{h}$$
$$\ell_1(x) = -\frac{x - x_i}{x_{i+1} - x_i} = \frac{1}{h}(x - x_i), \qquad \ell_1'(x_{i+1}) = \frac{1}{h},$$

por tanto

$$A_0(x_{i+2}) = 5,\ A_1(x_{i+2}) = -4,\ B_0(x_{i+2}) = 2h,\ B_1(x_2h) = 4h.$$

Así, resulta

$$y_{i+2} = 5y_i - 4y_{i+1} + h[2f_i + 4f_{i+1}].$$

♦

9.2.7. Análisis del error y de la estabilidad

En este apartado se estudia de modo muy somero los distintos tipos de error que acompañan a los métodos numéricos de integración de ecuaciones diferenciales.

Definición 9.2.1
Sea un método de integración numérico de ecuaciones diferenciales el cual proporciona una aproximación $\hat{y}(x_k + h)$ a partir los valores $y_0, y_1, \ldots, y_k$, y sea $y(x_k + h)$ el verdadero valor de la función y en el punto $x_k + h$. Llamamos **orden de truncamiento local** del método al orden en h de $\hat{y}(x_k + h) - y(x_k + h)$.

Definición 9.2.2
Diremos que un **método de integración** de ecuaciones diferenciales ordinarias es **consistente** si su orden de truncamiento local es mayor o igual que uno.

Para obtener dicho orden se procede a desarrollar en serie de Taylor $\hat{y}(x_k + h) - y(x_k + h)$ alrededor del punto x_0, de modo que diremos que el método tiene orden de truncamiento local n si se verifica $y(x_k + h) - \hat{y}(x_k + h) = O^{n+1}(h)$.

Desgraciadamente, cuando avanza la integración ya no se cuenta con los valores de la función que aparecen en el integrador, sino que únicamente se dispone de aproximaciones $\hat{y}(x_i)$, lo que induce nuevos errores.

Definición 9.2.3
Llamamos **error de truncamiento global** al error que aparece al aproximar $y(x_n)$ por $\hat{y}(x_n)$ siendo x_n el punto final de la integración, la cual se efectúa mediante el número de aplicaciones del método que resulte necesario.

El error de truncamiento global resulta un orden menor que el error de truncamiento local. En el cálculo de errores de truncamiento se supone que los cálculos se efectúan mediante una aritmética exacta, esto es, de infinitas cifras decimales.

Definición 9.2.4
Llamamos **errores de redondeo local y global** a los errores cometidos en cada aplicación del método y en el conjunto de ellas, debidos al número de cifras (finito) que se utiliza en los cálculos.

Definición 9.2.5
Llamamos **error global** del método al error total cometido al aproximar $y(x_n)$ por $\hat{y}(x_n)$, en el cual se acumulan todos los errores anteriores.

Para aproximar, mediante un método de pasos libres, la solución de una ecuación diferencial en un punto $x = b$ con error menor que un prefijado de antemano ϵ, se puede utilizar el siguiente método heurístico.

Sea la ecuación diferencial $y'(x) = f(x, y)$ con condición inicial $y(a) = y_0$. Para obtener $y(b)$ se utiliza un integrador numérico y se aproxima $y(b) \approx \hat{(y)}(x_0 + h)$ siendo $h = b - a$. A este valor se le denota como $y_1(b)$. A continuación, se divide el paso por dos $h = \frac{b-a}{2}$ obteniéndose mediante la aplicación del integrador la aproximación $y(b) \approx \hat{(}x_0 + 2h)$, la cual denotamos por $y_2(b)$. Dividiendo sucesivamente por dos el paso se obtiene la n-ésima aproximación $y(b) \approx \hat{y}(x_0 + 2^n h)$ donde $h = \frac{b-a}{2^{n-1}}$ la cual denotamos por $y_n(b)$. El proceso se repite hasta que $|y_k(b) - y_{k+1}(b)| \leq \epsilon$, tomándose en este momento $y_{k+1}(b)$ como valor de $y(b)$.

Ejemplo 9.2.15
Obtener $y(3)$ con error no superior a 1.10^{-3}, sabiendo que

$$y' = x + e^{-x} \operatorname{sen} y, \qquad y(1) = 1 \ .$$

Puesto que se piden tres cifras decimales exactas utilizaremos en nuestros cálculos cinco decimales como medida de protección.

En primer lugar, efectuaremos la integración en un solo paso $h = 2$, obteniéndose:

i	x_1	y_i	k_1	k_2	k_3	k_4	y_{i+1}
0	1	1	1.30956	2.10005	2.00562	2.95242	5.15778

por tanto, integrado en un paso (el número de pasos lo indicaremos como un superíndice entres paréntesis) se obtiene el valor $y_1(3) = 5.15778$.

Integrando en dos pasos ($h = 1$) se tiene la tabla:

i	x_1	y_i	k_1	k_2	k_3	k_4	y_{i+1}
0	1	1	1.30956	1.72234	1.71379	2.05615	2.70633
1	2	2.70633	2.05706	2.45411	2.44159	2.95486	5.17355

por tanto, integrado en dos pasos se obtiene $y_2(3) = 5.15355$, y por tanto, $||y_2(3) - y_1(3)|| = 0.016$ (con tres cifras decimales) es mayor que 10^{-3}.

Integrando en cuatro pasos ($h = 0.5$) se tiene la tabla:

i	x_1	y_i	k_1	k_2	k_3	k_4	y_{i+1}
0	1.0	1	1.30956	1.52806	1.53141	1.71891	1.76228
1	1.5	1.76228	1.71905	1.89130	1.88682	2.05714	2.70665
2	2.0	2.70665	2.05702	2.24165	2.23681	2.44816	3.82850
3	2.5	3.82850	2.44795	2.68842	2.68750	2.95539	5.17476

por tanto, integrado en dos pasos se obtiene $y_3(3) = 5.17476$, y con ello , $||y_3(3) - y_2(3)|| = 0.001$ (con tres cifras decimales) que es el error admisible, así podemos tomar $y(3) = 5.175 \pm 10^{-3}$.

♦

Definición 9.2.6
Sea la ecuación diferencial $y'(x) = f(x, y),\ x \in [a, b]$ con condición en $x_0 = a$ inicial $y(x_0) = y_0$. Diremos que un método de integración numérica $\hat{y}$ **es convergente** si $\forall x \in [a, b]$ se tiene que $\lim_{n\to\infty} \hat{y}(x_0 + nh) = y(x)$ donde $h = \frac{x-x_0}{n}$.

El problema de la convergencia es, en general, un problema difícil. En el caso de los métodos lineales multipaso este problema es más sencillo tal y como se verá.

Definición 9.2.7
Sea el método lineal multipaso dado por

$$a_k y_{k+i} + a_{k-1} y_{k+i-1} + \cdots + a_0 y_i = h(b_k f_{k+i} + \cdots + b_0 f_i).$$

Llamamos **primer y segundo polinomios de estabilidad** del método, a los polinomios $p(z)$ y $q(z)$ dados por

$$p(z) = a_k z^k + a_{k-1} z^{k-1} + \cdots + a_1 z + a_0$$
$$q(z) = b_k z^k + b_{k-1} z^{k-1} + \cdots + b_1 z + b_0.$$

Definición 9.2.8
Diremos que un método lineal multipaso es **estable** si las raíces r de $p(z) = 0$ satisfacen $|r| \leq 1$, y $|r| \leq 1$ caso de ser raíz múltiple.

A continuación se enuncian dos teoremas, los cuales no demostramos, que resuelven el problema de la convergencia para el caso de los métodos lineales multipaso.

Teorema 9.2.1
Sea un método lineal multipaso con primer y segundo polinomio de estabilidad dados por $p(z)$, $q(z)$. Entonces el método es consistente si, y sólo si, $p(1) = 0$ y $p'(1) = q(1)$.

Teorema 9.2.2
Un método lineal multipaso es convergente si, y sólo si, es estable y consistente.

9.3. Integración de sistemas y ecuaciones de orden n

Los métodos anteriores se generalizan con facilidad cuando se trata de resolver el problema:

$$\begin{aligned}
\dot{y}_1(x) &= f_1(y_1, y_2, \ldots, y_n, x)\\
\dot{y}_2(x) &= f_2(y_1, y_2, \ldots, y_n, x)\\
\cdots &= \cdots\\
\dot{y}_n(x) &= f_n(y_1, y_2, \ldots, y_n, x)\\
x \in [a, b] & \text{ con las condiciones iniciales}\\
y_1(a) &= y_{1,0}, \ldots, y_n(a) = y_{n,0}
\end{aligned}$$

para ello, basta tomar un método para una ecuación diferencial de primer orden y reemplazar y_i por $\overrightarrow{y}_i = (y_{1,1}, \ldots, y_{i,n})^t$, f por $\overrightarrow{f} = (f_1, \ldots, f_n)^t$ e y_0 por $\overrightarrow{y}_0 = (y_{0,1}, \ldots, y_{0,n})^t$, obteniéndose de este modo el método equivalente para sistemas.

Ejemplo 9.3.1
Escribir un método de Euler explícito para integrar el sistema

$$\begin{aligned}
\dot{y}_1 &= f_1(x, y_1, y_2)\\
\dot{y}_2 &= f_2(x, y_1, y_2)\\
x \in [x_0, x_0 + h],\ & y_1(x_0) = y_{1,0},\ y_2(x_0) = y_{2,0}.
\end{aligned}$$

El método de Euler explícito para el caso de una ecuación se escribe como

$$y_{i+1} = y_i + hf(x_i, y_i),$$

lo cual se generaliza como

$$\overrightarrow{y}_{i+1} = \overrightarrow{y}_i + h\overrightarrow{f}(x_i, \overrightarrow{y}_i),$$

lo que en forma matricial se expresa como

$$\begin{bmatrix} y_{1,i+1} \\ y_{2,i+1} \end{bmatrix} = \begin{bmatrix} y_{1,i} + hf_1(x, y_{1,i}, y_{2,i}) \\ y_{2,i} + hf_2(x, y_{1,i}, y_{2,i}) \end{bmatrix}.$$

♦

El problema de condición inicial para una ecuación de orden n viene dado por

$$\begin{aligned}
y^{n)}(x) &= f(x, y, y', y'', \ldots, y^{n-1)}),\ x \in [a, b]\\
y(a) = y_0,\ y'(a) &= y'_0,\ \ldots, y^{n-1)}(a) = y_0^{n-1)},
\end{aligned}$$

el cual se transforma en un sistema de ecuaciones diferenciales de orden n mediante el cambio de variable $z_i = y^{i)}$. Así, se tiene

$$\begin{aligned}
\dot{z}_1 &= z_2\\
\dot{z}_2 &= z_3\\
\cdots &= \cdots\\
\dot{z}_{n-1} &= z_n\\
\dot{z}_n &= f(x, z_1, z_2, \ldots, z_n)\\
\overrightarrow{z}_0 &= (y_{1,0}, y_{2,0}, \ldots, y_{n-1,0})^t,
\end{aligned}$$

el cual puede ser resuelto según las técnicas anteriores.

9.4. Introducción a los problemas de contorno para ecuaciones diferenciales ordinarias

En este capítulo se aborda el estudio de métodos numéricos para la resolución del problema de contorno para ecuaciones diferenciales ordinarias, esto es, problemas en los que en lugar de proporcionarse los valores de la función y sus derivadas en el extremo inicial del intervalo, se proporciona relaciones funcionales entre la función y sus derivadas en los extremos a y b; esto es, problemas de tipo

$$\begin{aligned}
&\frac{d\overrightarrow{y}}{dx} = \overrightarrow{f}(\overrightarrow{y}, x) \quad x \in]a, b[\\
&\phi_i(\overrightarrow{y}(a), \overrightarrow{y}(b)) = 0 \quad \forall i = 1, \ldots, n \\
&\overrightarrow{y} \in \mathbb{R}^n \quad \overrightarrow{f} : \mathbb{R}^n \to \mathbb{R}^n.
\end{aligned}$$

La solución de este problema es en general difícil, pero se simplifica notablemente cuando las condiciones de contorno son lineales en los valores de la función.

Método de tiro

El primer método que se aborda es el llamado método de tiro. Dicho método supone que el problema de condición inicial para nuestra ecuación se puede resolver, bien mediante integración exacta, bien mediante integración numérica. Sea pues $\overrightarrow{y}(x, \overrightarrow{y}_0)$ la solución del sistema de ecuaciones diferenciales dado anteriormente con la condición inicial $\overrightarrow{y}(a) = \overrightarrow{y}_0$. A partir de esta solución es posible determinar $\overrightarrow{y}(b) = \overrightarrow{y}(b, \overrightarrow{y}_0)$, y así las condiciones de contorno se pueden reescribir como:

$$\phi_i(\overrightarrow{y}_0, \overrightarrow{b}, \overrightarrow{y}_0) = \Psi_i(\overrightarrow{y}_0) = 0, \quad \forall i = 1, \ldots, n,$$

lo que representa un sistema de n ecuaciones en las incógnitas y_0^1,. . . ,y_0^n donde se ha utilizado la notación $\overrightarrow{y}_0 = (y_0^1, \ldots, y_0^n)$. Por tanto, el problema queda reducido a la resolución de un sistema de n ecuaciones con n incógnitas, lo cual se puede abordar a partir de un método de Newton-Ralphson (u otro método numérico adecuado). La resolución del sistema también puede abordarse a partir de métodos de optimización aplicados a la función $\Psi(y_0^1, \ldots, y_0^n) = \sum \text{lím}_{i=1}^2 \Psi_i^2(\overrightarrow{y}_0)$.

Ejemplo 9.4.1
Sea el problema de contorno dado por

$$y'' = -\frac{1}{2}y, \; x \in [0, \pi]$$

con las condiciones iniciales $y(0) = 1.0$, $y(1) = 0.5$. Resolver por el método de tiro utilizando como método de integración un método de Euler explícito con paso $\frac{1}{5}$.

En primer lugar, sea $y(x, \alpha)$ el valor que toma la función $y(x)$ cuando se integra con las condiciones iniciales $y(0) = 1$, $y'(0) = \alpha$ y sea $F(\alpha) = y(1, \alpha) - 0.5$.

La ecuación diferencial puede reducirse a un sistema mediante el cambio $z_1 = y$, $z_2 = y'$. Así, el problema resulta

$$\begin{aligned} z_1' &= \quad z_2 \\ z_2' &= -z_1. \end{aligned}$$

con la condición inicial $z_1(0) = 1$, $z_2(0) = \alpha$.

El método de Euler explícito resulta para esta ecuación

$$\begin{aligned} z_1(x_{i+1}) &= z_1(x_i) + h\, z_2(x_i) \\ z_2(x_{i+1}) &= z_2(x_i) - h\, z_1(x_i). \end{aligned}$$

Tomando inicialmente $\alpha = 0$, resulta la tabla de valores

x_i	0.0	0.4	0.6	0.8	1.0	
$z_1(x_i)$	1.0000	1.0000	0.9600	0.8800	0.7616	0.6080
$z_2(x_i)$	0.0000	-0.2000	-0.4000	-0.5920	-0.7680	-0.9203

por tanto, el valor de $F(0.0) = 0.6080 - 0.5000 = 0.1080$.

Para aproximar el valor de α para el cual $F(\alpha) = 0$ utilizaremos el método de Newton (podría haberse utilizado cualquier otro)

$$\alpha_{i+1} = \alpha_i - \frac{F(\alpha_i)}{F'(\alpha_i)}, \quad \alpha_0 = 0.$$

El valor $F'(0)$ lo aproximaremos por $F'(0) = \frac{F(0.01)-F(0)}{0.01}$, aproximación suficiente puesto que el método de Euler es de primer orden. Así, integrando las ecuaciones con condiciones iniciales $z_1(0) = 1$, $z_2(0) = 0.01$ se tiene

x_i	0.0	0.4	0.6	0.8	1.0	
$z_1(x_i)$	1.0000	1.0020	0.9640	0.8859	0.7693	0.6172
$z_2(x_i)$	0.0100	-0.1900	-0.3904	-0.5832	-0.7604	-0.9142

por tanto, el valor de $F(0.01) = 0.6172 - 0.5000 = 0.1172$, de donde $F'(0) = 0.9200$ y por tanto, $\alpha = -0.1174$.

Para evaluar $F(2.3360)$ es necesario integrar de nuevo con condiciones iniciales $z_1(0) = 1$, $z_2(0) = 2.3360$; así, se tiene

x_i	0.0	0.4	0.6	0.8	1.0	
$z_1(x_i)$	1.0000	0.97652	0.9130	0.8105	0.6714	0.5000
$z_2(x_i)$	-0.1174	-0.3174	-0.5127	-0.6953	-0.8574	-0.9917

Puesto que $F(-0.1174) = 0$, se tiene que la solución a nuestro problema viene dada por $y(x) = z_1(x)$, donde z_1 debe tomarse de la última tabla.

♦

Método de diferencias

El método de diferencias finitas consiste en introducir un mallado $x_0 = a < x_1 < x_2 < \cdots < x_n = b$ con $h_i = x_{i+1} - x_i$, $i = 0, \ldots, n-1$. Dado este mallado se procede a aproximar la ecuación diferencial y las condiciones de contorno mediante el uso de diferencias divididas; así, se tienen las aproximaciones:

$$\begin{aligned}
&f[x_{i-1}, x_i, xi+1] = p(x_i) f[x_{i-1}, x_{i+1}] + q(x_i) y(x_i) + r(x_i), \quad i = 1, \ldots, n-1 \\
&\alpha_1 y(x_0) + \beta_1 f[\ x_0\ , x_1\ ] = \gamma_1 \\
&\alpha_2 y(x_n) + \beta_2 f[x_{n-1}, x_n] = \gamma_2\ .
\end{aligned}$$

De este modo, se tiene un sistema de $n+1$ ecuaciones con $n+1$ incógnitas, lo que nos proporciona el valor de la función y en los nodos x_i.
La forma más habitual de discretizar el problema consiste en la introducción de un mallado regular $\omega = \{x_i = x_0 + ih | x_0 = 1, h = \frac{b-a}{n}\}$, y en aproximar las derivadas mediante operadores discretos adecuados, tales como los estudiados en el capítulo anterior.

Un esquema sencillo de primer orden para resolver el problema de contorno asociado a la ecuación lineal de segundo orden consiste en tomar las aproximaciones:

$$\begin{aligned}
y_i' &= \frac{y_{i+1} - y_{i-1}}{2h} + O(h^2), \quad \forall i = 1, \ldots, n-1 \\
y_i'' &= \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + O(h^2), \quad \forall i = 1, \ldots, n-1 \\
y_0' &= \frac{y_1 - y_0}{h} + O(h), \quad y_n' = \frac{y_n - y_{n-1}}{h} + O(h)
\end{aligned}$$

resultando el esquema de primer orden:

$$\begin{bmatrix}
b_1 & c_1 & 0 & 0 & \cdots & 0 & 0 & 0 \\
a_2 & b_2 & c_2 & 0 & \cdots & 0 & 0 & 0 \\
0 & a0_3 & b_3 & c_3 & \cdots & 0 & 0 & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & 0 & 0 & \cdots & a_{n-1} & b_{n-1} & c_{n-1} \\
0 & 0 & 0 & 0 & \cdots & 0 & a_n & b_n
\end{bmatrix}
\begin{bmatrix}
x_0 \\ x_1 \\ x_2 \\ \cdots \\ x_{n-1} \\ x_n
\end{bmatrix}
=
\begin{bmatrix}
h\gamma_1 \\ h^2 r_1 \\ h^2 r_2 \\ \cdots \\ h^2 r_{n-1} \\ h\gamma_2,
\end{bmatrix}$$

donde

$$\begin{aligned}
&b_1 = h\alpha_1 - \beta_1, \quad c_1 = \beta_1 \quad ; \qquad a_n = -\beta_2, \quad b_n = h\alpha_2 + \beta_2 \\
&a_i = 1 - \frac{h}{2} p_i, \quad b_i = -2 + h^2 q_i, \quad c_i = 1 + \frac{h}{2} p_i \quad \forall i = 2, \ldots n-1.
\end{aligned}$$

El sistema anterior resulta tridiagonal, por lo que puede ser fácilmente resuelto.

Para obtener un esquema de segundo orden basta sustituir en las condiciones de

contorno la aproximación a la primera derivada por:

$$
\begin{aligned}
y'_i &= \frac{y_{i+1} - y_{i-1}}{2h}, \quad i = 1, \ldots, n-1 \\
y'_0 &= \frac{1}{h}\left[\triangle - \frac{1}{2}\triangle^2\right] y_i = \frac{-3y_0 + 4y_1 - y_2}{2h} \\
y'_n &= \frac{1}{h}\left[\nabla + \frac{1}{2}\nabla^2\right] y_i = \frac{y_{n-2} - 4y_{n-1} + 3y_2}{2h},
\end{aligned}
$$

el sistema ahora no resulta tridiagonal (excepto si las condiciones de contorno se reducen a $y(a) = \gamma_1$, $y(b) = \gamma_2$), pero puede también ser reducido a tridiagonal mediante sencillas transformaciones.

Bibliografía

[1] ALDOUS J.M., WILSON R.J. (2000): *Graphs and Applications. An introductory approach.* Springer–Verlag, London.

[2] BAZARAA M.S., JARVIS. J.J. (1998): *Programación lineal y flujo de redes.* Limusa. México.

[3] BAZARAA M.S., SHERALI H.D., SHETTY C.M. (2006): *Non Linear Programing Theory and Alghorithms.* Jhon-Wiley&Sons, New York.

[4] BERGE, C. (1979): *Graphs and Hypergraphs.* North Holland, Amsterdam.

[5] BURDEN, R. B., FAIRES, D. J. (2002): *Análisis numérico.* Internacional Thompson Editores. México.

[6] CHAPRA, S.C., CANALE, R.P. (2003): *Métodos numéricos para ingenieros.* McGraw Hill.

[7] DEMIDOVITSCH B.P., MARON I.A. (1977): *Cálculo númerico fundamental.* Paranifo, Madrid.

[8] DEMIDOVITSCH B.P., MARON I.A., SCHUWALOWA E.S. (1980): *Métodos numéricos de análisis.* Paranifo, Madrid

[9] GERALD, C. - WHEATLEY, P. (2000): *Análisis numérico con aplicaciones.* Pearson Education, México.

[10] GIBBONS A.(1985): *Algorithmic Graph Theory.* Cambridge University Press, Cambridge.

[11] INFANTE R. (1997): *Programación matemática.* Ediciones de la U.N.E.D., Madrid.

[12] KINCAID, D., CHENEY, W. (1994): *Análisis numérico. Las matemáticas del cálculo científico.* Addison-Wesley Iberoamericana, Wilmington.

[13] *Kreyszig E. (2002)*: *Matemáticas avanzadas para ingeniería.* Limusa, México.

[14] LUENBERGER D.E. (1989): *Programción lineal y no lineal.* Addison-Wesley Iberoamericana, Wilmington,Delaware.

[15] WILSON R.J. (1983): *Introducción a la Teoría de Grafos.* Alianza Universidad, Madrid.